Contributions to posterior learning for likelihood-free Bayesian inference

Elouan Argouarc H

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2024IPPAS021

Thèse de doctorat

Contributions to posterior learning for likelihood-free Bayesian inference

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom SudParis

École doctorale n°574 École Doctorale de Mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques Appliquées

Thèse présentée et soutenue à Palaiseau, le 11/12/2024, par

ELOUAN ARGOUARC'H

Composition du Jury :

Mathilde MOUGEOT ENSIIE, ENS Paris-Saclay	Présidente
François SEPTIER Université Bretagne Sud	Rapporteur
Jean-Yves TOURNERET Université de Toulouse, ENSEEIHT-IRIT-TéSA	Rapporteur
Erwan LE PENNEC Institut Polytechnique de Paris, École Polytechnique	Examinateur
Guillaume CHARPIAT Université Paris-Saclay, INRIA	Examinateur
Eric BARAT Université Paris-Saclay, CEA	Invité
Eiji KAWASAKI Université Paris-Saclay, CEA	Invité
François DESBOUVRIES Institut Polytechnique de Paris, Télécom SudParis	Directeur de thèse

Remerciements

Je tiens tout d'abord à remercier les membres du Jury qui me font l'honneur d'assister à cette soutenance de thèse. Je remercie tout particulièrement les rapporteurs, Jean-Yves Tourneret et François Septier, pour leur lecture méticuleuse de ce manuscrit. Je tiens ensuite à remercier chaleureusement Eiji Kawasaki, Eric Barat et François Desbouvries qui m'ont accompagné durant cette thèse. Vous avez partagé avec moi votre expérience, vous m'avez fait bénéficier de votre expertise scientifique et technique, et vous m'avez accordé votre confiance dans ce projet de thèse.

Je tiens ensuite à exprimer ma profonde gratitude à tous mes proches, pour leur soutien indéfectible tout au long de ce parcours.

Merci à mon père, Emmanuel, et sa femme Sophie. Papa, merci d'être toujours là pour moi, de m'épauler au quotidien, et d'être un pilier dans les situations d'adversité. Tu ne ménages pas tes efforts lorsque j'en ai besoin, tu n'as reculé devant aucun sacrifice pour mes études, et je ne serai pas arrivé jusqu'ici sans toi. Merci Sophie pour ta bienveillance et ton affection inconditionnelles, elles m'apportent un soutien permanent.

Merci à ma mère, Nathalie, ton écoute et tes encouragements m'ont porté. Tu as toujours su trouver les mots justes pour m'accompagner dans les moments de doute et m'encourager à donner le meilleur de moi-même. Ton amour et ton dévouement m'inspirent chaque jour à avancer avec confiance.

Merci à ma sœur Sterenn, tu es constamment présente pour moi. Avec Romain, à qui j'adresse également de sincères remerciements, vous m'avez encouragé du début à la fin. Grâce à vous, j'ai la chance et l'immense joie de devenir Oncle Elou pour Paul et Azilis que j'aime de tout mon cœur.

Merci à mes grands-parents Tadig, Mammig et Soizig. Vous me tendez les bras toujours avec tendresse et la plus grande affection. Je vous remercie de m'avoir soutenu et encouragé. J'espère vous rendre fier en retour. Merci à Fabienne, ton soutien indéfectible est précieux.

Merci infiniment à mon oncle et ma tante, Sébastien et Béatrice. Je suis chanceux d'être toujours le bienvenu chez vous. J'arrive souvent un peu à l'improviste, mais toujours avec le sourire, car passer des moments avec vous est toujours synonyme de détente et de bonne humeur. Il y a parfois un parfum de vacances, parfois la convivialité des fêtes de Noël, mais venir chez vous à Brest a toujours la chaleur et le confort d'une deuxième maison. Votre soutien et votre affection inconditionnels m'accompagnent depuis toujours, et je ne vous remercierai jamais assez pour cela.

Merci à mes cousins Lisa, Eléonore et Théo. Théo, mon cousin, mon petit cousin avec qui je n'ai jamais ressenti la différence d'âge. On partage tellement de choses et ta maturité est impressionnante. Même s'il ne manque pas d'exemples de réussite dans la famille, je fais de mon mieux pour te donner le meilleur des exemples. Je te souhaite de réaliser ton potentiel, dans les études, dans le sport, et je serai là pour t'épauler, comme tu l'as été pour moi.

Eléonore, merci pour ton soutien moral. Tu m'apportes de la bonne humeur en permanence

et nous partageons tellement de bons moments. Nous avons tellement de bons souvenirs, des fous-rires, des frissons aussi.

Lisa. Je ne saurai pas quoi faire sans toi. Tu es la première personne vers laquelle je me tourne, dans n'importe quelle situation. Tu m'apportes le soutien moral dont j'ai besoin et tu sais toujours trouver les mots justes. Je suis tellement reconnaissant et heureux de te savoir à mes côtés.

Il y a un théorème qui stipule qu'un tabouret est stable lorsqu'il dispose de trois pieds non-alignés. Merci donc à mes amis les plus proches, Akutino, Mathis et Paul. Je suis content de savoir qu'on peut compter les uns sur les autres, depuis de nombreuses années. Aku, tu es l'un de mes amis de plus longue date. Tu es à l'écoute, constamment prêt à rendre service et toujours disponible pour les autres. C'est vers toi que je me tourne pour vider mon sac, me confier. Je sais que je peux compter sur ton soutien quoiqu'il arrive, et que jamais tu ne me laisserai dormir sous un pont. Tu m'as constamment épaulé durant cette période de thèse. Avant même le premier jour, tu m'écoutais parler du sujet, tu m'as accompagné pour les entretiens. Logiquement, tu restes mon supporter numéro 1 le dernier jour: outre gérer les urgences vestimentaires, tes encouragements m'ont aidé à aborder l'échéance de cette soutenance avec plus de sérénité. Je compte sur ton soutien pour les prochaines étapes de nos vies, et tu peux compter sur le mien. Enfin, je suis ravi que tu fasses partie de la famille, et je suis fier que ce soit en partie grâce à moi. Mathis, tes messages, presque au quotidien, me rappellent que tu m'épaules. On se soutient mutuellement depuis que l'on se connaît: le bac, la prépa, les stages, les premiers jobs. Même à distance, même à l'international, tu m'apportes ton soutien et je suis content que tu reviennes souvent nous voir. Ton parcours est un exemple de réussite: tu as dit que tu irais là où tu es, et que tu y ferais ce que tu fais. Félicitations, mais ce n'est que le début. Paul, l'un des piliers de cette aventure. Surtout en dernière année. Non seulement tu as été présent pour dynamiser les week-ends et les vacances ce qui a permis de me sortir du quotidien, mais tu as aussi été à mon écoute en permanence, en faisant preuve de bienveillance et de soutien. S'il y en a un qui ne me laisserait jamais seul, derrière, c'est bien toi.

Mille mercis à mes colocataires Sasila et Fabiola. Aucune coloc n'égalera la nôtre. Nous avions nos habitudes et nos rendez-vous, un quotidien agréable et de nombreuses occasions de sorties ensemble. Parfois des discussions animées, des potins et du drama pour casser la routine. Sasila, j'envierai parfois presque un retour du confinement tellement j'ai un bon souvenir de la colocation à Choisy ensemble. Merci pour la bonne humeur au quotidien, les fous rires, les séances de sports, les plats de gourmets que tu préparais, et pour tout le reste. Fabiola, merci pour ton soutien inconditionnel. Tu as été présente pour m'épauler pendant 3 années de thèse, de Choisy jusqu'à Massy, du début jusqu'à la fin. Nous avons partagé tant de moments, les hauts, les bas, les réussites, certains de nos échecs. Je suis fier de vos parcours, ce n'est que le début d'une vie personnelle et professionnelle épanouie, je vous le souhaite.

Merci Yazid. Tu me réponds constamment, parfois à pas d'heures, pour discuter, écrire, corriger, brainstormer, coder, débuger ensemble. Ta réussite académique est une source de motivation et je ne désespère pas que l'on puisse collaborer un jour pour écrire un papier comme on se l'était promis. Tu es aussi le premier de mes amis à t'être marié, j'étais très heureux d'être convié pour partager ce moment avec toi et Soukaina. Enfin, merci de m'avoir fait découvrir Gemüse et Surpriz.

Je remercie les Couz Camille, Laurianne, Morgane et Tyfène. Merci à Karl. Merci à mes amis de Télécom SudParis Julien, Mahdi, Farouk, Youssef, Ramzi et Ryan, les charos. Merci à Baptiste et Steven, mes piliers de classe prépa. Merci à Quentin.

Enfin, un immense merci à Cristiana. Sans toi, tout ceci n'aurait que peu de saveurs.

Introduction

Contexte Général

L'inférence bayésienne a posteriori est une méthodologie générale qui, une fois la valeur d'une observation Y donnée, permet de découvrir les valeurs probables prises par une quantité inconnue d'intérêt X (appelée *label*) liée à Y, décrite par la distribution de probabilité a posteriori $(X|Y)$. Cette méthodologie peut être appliquée lorsque le praticien est capable (i) de spécifier (ou d'assumer un modèle pour) la distribution conjointe du couple de variables (X, Y) et (ii) de calculer, ou d'estimer via l'échantillonnage de Monte Carlo, la distribution de probabilité a posteriori.

Une distribution de probabilité conjointe sur (X, Y) peut en effet être factorisée en $(Y|X)$ multiplié par (X). Par conséquent, en pratique, spécifier la distribution de probabilité conjointe peut se faire en spécifiant deux distributions distinctes : (i) un modèle d'observation qui est une distribution conditionnelle sur $(Y|X)$ et (ii) une distribution a priori sur (X). Faire des hypothèses pertinentes sur ces deux distributions de probabilité par rapport au problème à traiter est effectivement une exigence majeure pour obtenir des informations pertinentes dans l'inférence et est un sujet largement couvert dans la littérature. Cependant, le problème de l'obtention de telles distributions n'est pas celui que nous traitons.

Dans cette thèse, nous supposons plutôt que nous disposons d'une distribution a priori pertinente et d'un modèle d'observation précis, mais que ce dernier possède une fonction de densité de probabilité (PDF) qui n'est pas calculable, rendant les PDF jointes et a posteriori inutilisables en pratique. Nous supposons plutôt que le modèle d'observation est partiellement connu via un jeu de données composé de couples enregistrés, tels que les (x_i, y_i) qui le composent sont effectivement liés par la distribution conditionnelle correspondante. Notre objectif est donc d'exploiter cet ensemble de données et de construire un modèle qui capture la dépendance entre les deux variables aléatoires et, en fin de compte, permette d'approximativer la distribution a posteriori d'intérêt.

Cette formulation générale inclut les tâches usuelles d'apprentissage statistique de classification et de régression, qui ont suscité un intérêt croissant à l'ère du big data et avec le développement des méthodes d'apprentissage (profond). Cette formulation propose également une méthodologie alternative aux méthodes ABC, dans le contexte du problème de l'inférence basée sur la simulation, qui est devenue populaire récemment dans de nombreux domaines scientifiques car elle permet de confronter des mesures du monde réel à des modèles in vitro ou in silico conçus comme des modèles de simulation.

La formulation bayésienne du problème, qui décrit les valeurs probables de X as-

socié à une observation Y en utilisant la distribution de probabilité a posteriori, permet de prendre en compte de manière directe l'incertitude aléatoire. Cependant, lorsqu'on utilise un modèle pour la distribution de probabilité a posteriori inconnue en utilisant un ensemble de données qui résume le modèle d'observation indisponible, il est également souhaitable de prendre en compte l'incertitude épistémique de modélisation, qui est l'un des problèmes considérés dans cette thèse. Cette thèse comprend plusieurs contributions méthodologiques qui, ensemble, peuvent aider à comparer différentes approches d'approximation a posteriori sous l'angle de l'échantillonnage de Monte Carlo et de la quantification de l'incertitude.

Contenu du document

Chapitre 1/

Dans le chapitre 1, nous proposons une visite guidée depuis les méthodes d'inférence bayésienne classiques jusqu'à l'apprentissage statistique d'une distribution a posteriori, qui sert d'introduction générale aux problèmes considérés dans cette thèse. Nous expliquons d'abord que l'inférence bayésienne a posteriori, avec des hypothèses appropriées sur les distributions de probabilité a priori et du modèle d'observation, permet l'étude des processus réels. Cela se fait cependant au prix de plusieurs difficultés méthodologiques et computationnelles, telles que l'estimation des quantités statistiques d'intérêt à l'aide de l'échantillonnage de Monte Carlo.

Nous considérons ensuite le cas où nous disposons d'un modèle d'observation pertinent, mais dont la PDF n'est pas calculable, et cette distribution de probabilité est seulement connue via des observations collectées en un jeu de données. Cette situation se produit, par exemple, lorsque le modèle d'observation est défini comme un simulateur ou lorsque le modèle d'observation n'est plus disponible. Dans ce cas, la vraisemblance et la PDF a posteriori deviennent des fonctions inutilisables en l'état.

Nous en arrivons alors au sujet de cette thèse. Nous expliquons que l'approche d'apprentissage statistique, qui consiste à utiliser des couples produits par le modèle d'observation pour obtenir une estimation de l'a posteriori, fournit effectivement une solution méthodologique possible. Enfin, nous présentons les différents défis associés à l'apprentissage statistique d'une distribution a posteriori, dont certains sont abordés dans cette thèse.

Chapitre 2/

Dans le chapitre 2, nous nous concentrons sur la méthode d'estimation de la distribution a posteriori en utilisant l'approximation du *likelihood-to-evidence-ratio*. Nous pouvons comprendre cette méthode comme une approximation paramétrique d'une distribution d'intérêt—dans notre cas, la distribution a posteriori—en utilisant un modèle non normalisé spécifique basé sur un classifieur et une distribution instrumentale—dans notre cas, la distribution a priori.

Dans ce travail, nous nous concentrons sur le problème de l'échantillonnage à partir

de l'approximation correspondante. Le point précis de notre contribution consiste à exploiter la structure sous-jacente afin d'obtenir différentes procédures d'échantillonnage de la distribution de probabilité correspondante qui soient faciles à mettre en œuvre. Nous proposons d'effectuer un échantillonnage approximatif d'une distribution cible en utilisant une distribution instrumentale avec les techniques classiques d'échantillonnage de Monte Carlo basées sur le rapport de PDF, mais où ce rapport est remplacé par une approximation basée sur un classifieur (voir section 2.1). Cette méthode d'échantillonnage approximatif peut alors être comprise comme un échantillonnage exact, mais où la distribution échantillonnée est une approximation non normalisée basée sur un classifieur et une distribution instrumentale. L'intérêt de notre méthodologie est au moins double: l'algorithme d'échantillonnage résultant est sans paramètre et peut être appliqué dans le contexte où la distribution a priori (instrumentale) possède elle aussi une PDF incalculable.

Enfin, nous développons davantage sur la connexion entre notre approche d'échantillonnage approximatif et les modèles à énergie (voir section 2.2). Nous proposons enfin une autre application de la méthodologie du classifieur qui approxime un rapport de PDF, dans le contexte de la modélisation générative. En effet, nous proposons un moyen efficace et pratique de pallier la contrainte de bijectivité des flots normalisants (voir section 2.3).

Chapitre 3/

Dans le chapitre 3, nous abordons le problème de la quantification de l'incertitude (UQ), qui est une préoccupation cruciale lorsque la quantité de données disponibles est limitée. Ce sujet est largement couvert dans la littérature historique et récente. Dans ce contexte, les défis actuels visent à adapter la quantification bayésienne de l'incertitude à des réseaux de neurones profonds à grande échelle.

Dans ce travail, à proprement parler, nous ne proposons pas de nouvelle méthode pour quantifier l'incertitude. Nous effectuons plutôt une comparaison entre les approches de modélisation générative et discriminative dans le cadre de la quantification bayésienne de l'incertitude. En analysant la distribution prédictive a posteriori (PPD), nous expliquons que, bien que leurs constructions soient très similaires, dans les modèles génératifs et discriminatifs, les objets interagissent de manières très différentes, entraînant des différences structurelles entre les deux approches. Plus particulièrement, nous analysons le rôle d'une distribution a priori, qui est explicite dans le cas génératif et implicite dans le cas discriminatif; nous analysons cette différence et fournissons des informations sur pourquoi les modèles discriminatifs souffrent de déséquilibres dans les ensembles de données. Nous analysons également le rôle des observations dans l'inférence globale : il est double, lié à la fois aux incertitudes aléatoires et épistémiques dans le cas de la modélisation générative, alors qu'il est uniquement lié à l'incertitude épistémique dans la modélisation discriminative. Nous tirons parti de cette conclusion pour expliquer la différence intrinsèque entre la PPD dans les deux cas, pour laquelle nous proposons un schéma général d'échantillonnage de Gibbs, et l'incompatibilité structurelle de l'approche discriminative avec la tâche d'apprentissage semi-supervisé.

Chapter 4/

Dans le chapitre 4, nous abordons le problème de l'approximation variationnelle flexible en utilisant un modèle avec une PDF calculable. Ce problème s'étend naturellement des chapitres précédents, car les modélisations génératives et discriminatives nécessitent toutes deux des PDF tractables des modèles sous-jacents. Avoir une PDF explicite et tractable est (i) avantageux sur le plan computationnel, car cela permet l'utilisation de diverses techniques de Monte Carlo, et (ii) une exigence pour une quantification bayésienne de l'incertitude (UQ) exacte.

Nous commençons par examiner les méthodes conventionnelles de construction de distributions de probabilité paramétriques, en nous concentrant sur les mécanismes qui produisent des PDF tractables (voir section 4.1). Nous passons également en revue brièvement le sujet de la reparamétrisation des gradients, car c'est un problème pertinent dans le contexte de ce travail (voir section 4.2). Ensuite, dans la section 4.3, nous introduisons *Discretely Indexed Flows* (DIF), un nouveau modèle paramétrique qui étend les modèles de mélange de gaussiennes. Dans ce modèle, nous remplaçons les poids de mélange constants par une fonction de réseau de neurones, plus précisément un classifieur, pour améliorer la flexibilité et l'expressivité dans les problèmes d'approximation variationnelle. Les DIF offrent plusieurs avantages : (i) une évaluation rapide et exacte de la PDF, (ii) un schéma d'échantillonnage simple, et (iii) une approche de reparamétrisation des gradients qui le rend adapté à divers types de problèmes d'approximation variationnelle. Enfin, nous expliquons comment appliquer cette construction à l'inférence variationnelle, à l'estimation de densité et à l'estimation de densité conditionnelle.

La trame narrative

Comme nous venons de le voir, les différents chapitres de cette thèse (et leurs contributions sous forme d'articles correspondants - voir la section suivante) sont présentés de manière relativement indépendante les uns des autres. À première vue, ils peuvent sembler aborder des problèmes spécifiques et sans rapport, mais en essence, ils sont liés par un fil conducteur commun. En particulier, en synthétisant et en reliant les arguments et conclusions de chaque chapitre, nous obtenons un récit cohérent qui fournit des perspectives interconnectées et met en évidence les thèmes récurrents.

Nous explicitons maintenant ce récit sous la forme des trois phrases suivantes, qui établissent en effet des liens entre les différents chapitres, et nous développons brièvement ces phrases afin de clarifier le fil conducteur commun.

- **"La méthode du likelihood-to-evidence-ratio..."**

  L'approximation de la PDF a posteriori utilisant le *likelihood-to-evidence-ratio* est un outil très utile. Il a récemment suscité un grand intérêt, car il permet de transformer un problème de l'estimation de densité conditionnelle en un problème de classification binaire. Dans la première contribution de cette thèse, nous avons abordé le problème de l'échantillonnage à partir de l'approximation a posteriori correspondante en utilisant les algorithmes d'échantillonnage basés sur le rapport

habituel.

"...correspond à une modélisation avec PDF non-normalisée..."

Cependant, cette approximation est intrinsèquement un modèle à énergie, et en tant que tel, PDF n'est disponible que sous une forme non normalisée. Bien que la constante de normalisation intraitable associée soit indépendante du label d'intérêt dans l'inférence a posteriori, elle dépend des paramètres du modèle.

"...et n'est pas directement compatible avec la tâche de quantification d'incertitude."

La quantification de l'incertitude épistémique est devenue un sujet prévalent dans l'apprentissage statistique moderne, particulièrement à l'ère de l'apprentissage machine (profond). Les méthodes bayésiennes pour la quantification d'incertitude considèrent les paramètres comme des variables aléatoires et se concentrent sur l'échantillonnage à partir de la distribution PPD. Pour certains modèles non normalisés, la manière appropriée de réaliser l'UQ épistémique reste floue et peut constituer un sujet de recherche futur.

Cela nous conduit à considérer spécifiquement des constructions qui sont effectivement compatibles avec la tâche de l'UQ épistémique. Cela introduit donc le Chapitre 3, dans lequel nous comparons les modèles génératifs et discriminatifs, qui sont en effet compatibles avec ce problème.

- **"Les constructions génératives et discriminantes..."**

  Les modèles génératifs et discriminatifs exploitent tous deux une distribution de probabilité conditionnelle paramétrée, mais diffèrent dans leur construction. La première modélisation paramètrise la distribution des observations conditionnallement étiquettes, tandis que le second fait le contraire. Cette différence entraîne de nombreuses conséquences intéressantes;

  "...permettent la quantification de l'incertitude épistémique,..."

  Sous une hypothèse cruciale mentionnée ci-après, les approches génératives et discriminantes permettent toutes deux d'échantillonner suivant la PPD. En effet, d'une part, les modèles discriminatifs offrent une approche directe pour échantillonner la PPD et pour réaliser une inférence prenant en compte l'incertitude. D'autre part, bien que l'approche de modélisation générative, comme la modélisation basée sur le likelihood-to-evidence-ratio, fournisse une approximation non normalisée de la PDF a posteriori, nous expliquons comment il est tout de même possible de réaliser une quantification épistémique de l'incertitude via la PPD. Dans ce contexte, nous avons réalisé, dans la deuxième contribution de cette thèse, une comparaison approfondie des approches génératives et discriminatives à travers une

analyse de la PPD, et avons conclu sur le comportement des deux approches dans le cadre de l'apprentissage semi-supervisé et des jeux de données déséquilibrés. Nous avons également proposé une méthode pratique pour réaliser une inférence tenant compte de l'incertitude pour les modèles génératifs via l'échantillonnage de la PPD.

"...à condition que ces constructions utilisent un modèle a PDF tractable."

L'hypothèse qui permet la quantification l'incertitude épistémique pour les approches génératives et discriminatives est celle d'un modèle paramétrique conditionnel bénéficiant d'une fonction PDF normalisée et tractable. Plus précisément, cette hypothèse produit un problème d'inférence qui permet l'échantillonnage de la PPD. Cela nous conduit donc du Chapitre 3, où nous abordons le problème de l'UQ épistémique avec les approches discriminatives et génératives, au Chapitre 4, qui traite du problème de la construction de modèles paramétriques avec des PDF tractables. Cela nous mène directement à la question suivante :

- **"Comment construire un modèle paramétrique (possiblement conditionnel) qui bénéficie d'une PDF calculabe ?"**

Cette question est précisément le sujet abordé dans le Chapitre 4. Un modèle avec une PDF tractable est en effet un outil puissant dans l'apprentissage statistique paramétrique, car il peut être utilisé pour l'estimation de densité, l'estimation de densité conditionnelle dans la modélisation générative et discriminative d'une PDF a posteriori, l'inférence variationnelle, et l'échantillonnage approximatif à partir d'une distribution connue par des échantillons enregistrés. Cependant, les modèles de distributions de probabilité flexibles avec des PDF tractables sont quelque peu limitées aux méthodes usuelles de (i) changement de variables avec des flots normalisants, et (ii) modèles de variables latentes discrètes tels que les mélanges. Dans la troisième contribution de cette thèse, nous proposons une nouvelle construction qui combine effectivement ces deux mécanismes, et bénéficie d'une flexibilité variationnelle accrue (en utilisant des fonctions de réseaux de neurones) tout en conservant une évaluation exacte de la PDF simple. Cette construction peut facilement être transformée en modèles conditionnels pour la modélisation générative et discriminative d'une approximation d'une PDF a posteriori.

Contributions

Cette thèse est basée sur les articles de journaux suivants. Nous décrivons la correspondance entre les articles et les sections de cette thèse dans la table suivante 1.

- "Binary Classification Based Monte Carlo Simulation", **Elouan ARGOUARC'H**, François DESBOUVRIES; publié dans IEEE Signal Processing Letters, vol. 31, pp. 1449-1453, 2024, doi: 10.1109/LSP.2024.3396403.

Chapitre 2	"Binary Classification based Monte Carlo sampling"	Section 2.1
Chapitre 3	"Generative vs. Discriminative Bayesian Posterior learning"	-
Chapitre 4	"Discretely Indexed Flows"	Section 4.3

Table 1: Correspondance entre papiers et sections

https://ieeexplore.ieee.org/document/10517652;

- "Generative vs. Discriminative Bayesian Posterior learning", **Elouan ARGOUARC'H**, François DESBOUVRIES, Eric BARAT, Eiji KAWASAKI; soumis à *Bayesian Analysis*. Disponible sur ArXiv:stat/2406.09172. - https://arxiv.org/abs/2406.09172;
- "Discretely Indexed Flows", **Elouan ARGOUARC'H**, François DESBOUVRIES, Eric BARAT, Eiji KAWASAKI, Thomas DAUTREMER; prépublication. Disponible sur ArXiv:stat/2204.01361. - https://arxiv.org/abs/2204.01361.

En plus de ces trois articles, des versions préliminaires ont également été publiées dans une conférence nationale de traitement du signal et de l'image. Elles sont liées à cette thèse, respectivement dans les parties du chapitre 4 et du chapitre 3.

- "Flots stochastiques discrets" **Elouan ARGOUARC'H**, François DESBOUVRIES, Eric BARAT, Eiji KAWASAKI, Thomas DAUTREMER; Actes du 28ème colloque GRETSI, Nancy, France, September 2022.

  https://www.gretsi.fr/data/colloque/pdf/2022_argouarch1049.pdf;

- "Apprentissage Bayésien Semi-supervisé par modélisation générative", **Elouan ARGOUARC'H**, François DESBOUVRIES, Eric BARAT, Eiji KAWASAKI, Thomas DAUTREMER; Actes du 29ème colloque GRETSI,Grenoble, France, September 2023.

  https://www.gretsi.fr/data/colloque/pdf/2023_argouarch1365.pdf.

Introduction

General Context

Bayesian Posterior inference is a general methodology that, once given the value of an observation Y, enables the unraveling of the probable values taken by an unknown quantity of interest X related to Y, encapsulated in the posterior probability distribution $(X|Y)$. This methodology can be applied when the practitioner is able to (i) specify (or assume a model for) the joint distribution of the pair of variables (X, Y) and (ii) compute, or estimate via Monte Carlo sampling, the posterior probability distribution.

A joint probability distribution over (X, Y) can indeed be factorized as $(Y|X)$ times (X). Therefore, in practice, specifying the joint probability distribution can be fulfilled by specifying two distinct distributions: (i) an *observation model* which is a conditional distribution over $(Y|X)$ and (ii) a *prior* which is a distribution over (X). Making relevant assumptions about these two probability distributions in regard to the problem at hand is indeed a main requirement for obtaining relevant information in the inference and is a widely covered topic in the literature. However, the problem of eliciting such distributions is not the problem that we aim to tackle.

In this thesis, we instead suppose that we are given a relevant prior distribution and an accurate observation model, but the latter has an intractable probability density function (PDF), making both the joint and the posterior PDFs unavailable. We instead suppose that the observation model is partially known via a dataset composed of recorded couples, which are indeed related by the corresponding conditional distribution. Our goal is therefore to leverage this dataset and build a model that captures the dependency between the two random variables (RVs) and ultimately approximates the unavailable posterior distribution of interest.

This general formulation includes the statistical learning tasks of classification and regression, which have gained increasing interest in the big-data era and with the development of (deep) machine-learning methods. This formulation also proposes an alternative methodology to Approximate Bayesian Computation methods, in the context of Simulation Based Inference, which has recently become popular in many scientific fields, since they enable to confront real world measurements to intrinsic *in vitro* or *in silico* observations models designed as simulation models.

The bayesian formulation of the problem, which describes probable values X given an observed Y using the posterior probability distribution, provides with a straightforward accounting of the *aleatoric* uncertainty. However, when using a model for the unknown posterior probability distribution using a dataset which summarizes the unavailable

observation model, it is also desirable to account for the *epistemic* modeling uncertainty, which is one of the problems considered in this thesis. This thesis comprises several methodological contributions which, together, can help comparing different posterior approximation approaches under the scope of Monte-Carlo sampling and uncertainty quantification (UQ).

Contents of this thesis

Chapter 1/

In chapter 1, we propose a guided tour from classical Bayesian inference to posterior statistical learning, which serves as a general introduction to the problems considered in this thesis. We first explain that Bayesian posterior inference, with appropriate assumptions of prior and observation model probability distributions, enables the study of real-world processes. This comes, however, at the cost of several methodological and computational shortcomings, such as that of estimating statistical quantities of interest using Monte Carlo sampling.

We then consider the case where we dispose of a given relevant observation model but which has an intractable PDF and is instead only known via recorded observations. This situation occurs, for instance, when the observation model is defined as a simulator or when the observation model is no longer available. In this case, both the likelihood and the posterior PDF become unavailable functions.

We then come to the topic of this thesis. We explain that the statistical learning approach, which consists in using recorded samples from the observation model to obtain an estimate of the posterior, indeed provides a solution to this shortcoming. We finally present different challenges associated with statistical learning of a posterior distribution, some of which are tackled in this thesis.

Chapter 2/

In chapter 2, we focus on the method for estimating the unavailable posterior using the likelihood-to-evidence ratio approximation. We can understand this method as a parametric approximation of a distribution of interest—in our case, the posterior—using a specific unnormalized model based on a classifier ratio and an instrumental distribution—in our case, the prior.

In this work, we focus on the problem of sampling from the corresponding approximation of the posterior distribution. The score point of our contribution is to leverage the underlying structure in order to obtain different easy-to-carry-out sampling procedures from the corresponding probability distribution. We propose to perform approximate sampling from a target distribution using an instrumental distribution with the classical ratio-based Monte Carlo sampling techniques, but where the PDF ratio is replaced by a classifier-ratio approximation (see section 2.1). This method of approximate sampling can then be understood as exact sampling from an unnormalized classifier-ratio based approximation. The interest of our methodology is at least twofold: the resulting sam-

pling algorithm is parameter-free and can be applied in the context where the prior distribution has an intractable PDF.

Additionally, we elaborate on the connection between our approximate sampling approach and energy-based modeling (see section 2.2). We finally propose another application of the classifier-ratio methodology in the context of generative modeling with an efficient and practical way to cope with the bijectivity constraint of Normalizing Flows (see section 2.3).

Chapter 3/

In chapter 3, we aim to tackle the problem of UQ, which is a crucial concern when the amount of available data is limited. This topic is widely covered in both historical and recent literature. In this context, current challenges aim to scale Bayesian UQ to large-scale deep-neural networks.

In this work, properly speaking, we do not propose a novel method for quantifying the uncertainty. We instead carry out a comparison between the generative and discriminative modeling approaches under the scope of Bayesian UQ. By analyzing the posterior predictive distribution (PPD), we explain that, even though their constructions are very similar, in generative and discriminative models, the objects interact in very distinct ways, leading to structural differences between the two approaches. Most notably, we analyze the role of a prior distribution, which is explicit in the generative case and implicit in the discriminative case; we analyze this difference and provide insights into why discriminative models suffer from imbalanced datasets. We also analyze the role of observations in the global inference: it is dual, both related to aleatoric and epistemic uncertainties in the case of generative modeling, while only related to epistemic one in discriminative modeling. We leverage this conclusion to explain the intrinsic difference between the PPD in both cases, for which we propose a general Gibbs sampling scheme, and the structural incompatibility of the discriminative approach with the task of semi-supervised learning.

Chapter 4/

In this chapter, we address the problem of flexible variational approximation using a model with a tractable PDF. This problem naturally extends from the previous chapter, as both generative and discriminative modeling require tractable PDFs of the underlying models. Having an explicit and tractable PDF is (i) computationally convenient as it enables the use of various Monte Carlo techniques, and (ii) a requirement for exact Bayesian UQ.

We begin by reviewing conventional methods for constructing parametric probability distributions, focusing on mechanisms that yield tractable PDFs (see section 4.1). We also briefly review the topic of reparameterization of gradients since it is a relevant problem in the context of this work (see section 4.2). Next, in section 4.3, we introduce Discretely Indexed Flows (DIF), a novel parametric model that extends Gaussian Mixture Models. In this model, we replace constant mixture weights with a classifying neural network function to enhance flexibility and expressiveness in variational approx-

imation problems. DIF offers several advantages: (i) rapid and exact PDF evaluation, (ii) a straightforward sampling scheme, and (iii) a gradient reparameterization approach that renders it suitable for various types of variational approximation problems. Finally, we explain how to apply this construction to variational inference, density estimation, and conditional density estimation.

The narrative of this thesis

As we have just seen, the different chapters of this thesis (and their corresponding article contributions, see next section) are presented rather independently of one another. At first glance, they might seem to tackle unrelated and specific problems, but in essence, they are linked by a common thread. In particular, by synthesizing and bridging arguments and conclusions from each chapter, we obtain a cohesive narrative, which provides interconnected insights and highlights the overlapping themes.

We now explicit this narrative in the form of the three following sentences, which indeed establish connections between the different chapters, and we briefly elaborate on these sentences in order to clarify the common narrative.

- **"The likelihood to evidence ratio..."**

  The likelihood-to-evidence ratio is a very useful tool for approximating an unavailable posterior PDF. It has gained interest in recent years as it enables turning the problem of conditional density estimation into a problem of binary classification. In the first contribution of this thesis, we have tackled the problem of sampling from the corresponding posterior approximation by using the usual ratio-based sampling algorithms;

  "...corresponds to an underlying unnormalized model..."

  However, the likelihood-to-evidence ratio approximation is inherently an energy-based model, and as such, its PDF is only available in an unnormalized form. While its associated intractable normalizing constant is independent of the label variable of interest in the posterior inference, it does depend on the model parameters;

  "...and is not compatible with the task of epistemic uncertainty quantification."

  Epistemic UQ has become a prevalent topic in modern statistical learning, especially in the (deep) machine learning era. Bayesian methods for epistemic UQ consider parameters as RVs and focus on sampling from the PPD. For some unnormalized models, particularly likelihood-to-evidence ratio-based models, the appropriate way to perform epistemic UQ remains unclear and can be a topic for future work. This leads us to specifically consider constructions that are indeed compatible with the task of epistemic UQ. This therefore introduces Chapter 3,

in which we compare the generative and discriminative models, that are indeed compatible with this problem.

- **"Generative and discriminative models..."**

  Generative and discriminative models both leverage a parameterized conditional probability distribution but differ in their construction: the former parameterizes the distribution of observations given labels, while the latter does the opposite. This difference has many interesting consequences;

  "...indeed enable epistemic uncertainty quantification,..."

  Under a crucial assumption mentioned hereafter, both generative and discriminative approaches enable sampling from the PPD. Indeed, on the one hand, discriminative models provide a straightforward approach for sampling from the PPD and performing uncertainty-aware inference. On the other hand, although the generative modeling approach, like likelihood-to-evidence ratio modeling, provides an unnormalized approximation of the posterior PDF, we explain how it is still possible to perform epistemic UQ via the PPD. In this context, we conducted, in the second contribution of this thesis, a thorough comparison of generative and discriminative approaches through an analysis of the PPD, and concluded on the behavior of both approaches under the scope of semi-supervised learning and unbalanced datasets. We also provided a practical way to perform uncertainty-aware inference for generative models;

  "...under the condition that both approaches use a parametric model with tractable and normalized PDF."

  The assumption that enables epistemic UQ for both generative and discriminative approaches is that of a conditional parametric model which benefits from a normalized and tractable PDF. More precisely, this assumption yields a tractable inference problem, and enables sampling from the PPD. This therefore leads us from Chapter 3, in which we tackle the problem of epistemic UQ with discriminative and generative approaches, to Chapter 4, which tackles the problem of constructing parametric models with tractable PDFs. This directly leads us to the next question:

- **"How do we build a (possibly conditional) parametric model which benefits from straightforward PDF evaluation ?"**

  This question is precisely the topic covered in Chapter 4. A model with a tractable PDF is indeed a powerful tool in parametric statistical learning, as it can be used for density estimation, conditional density estimation in generative and discriminative modeling of a posterior PDF, variational inference, and approximate sampling from a distribution known by recorded samples. However, flexible probability distributions with tractable PDFs are somewhat limited to the usual methods of (i)

change of variables with normalizing flows, and (ii) discrete latent variable models with mixture models. In the third contribution of this thesis, we propose a new construction which indeed combines these two mechanisms, and benefits from increased variational flexibility (using NN functions) and retain straightforward exact PDF evaluation. This construction can easily be turned into conditional models for generative and discriminative modeling of a posterior approximation in this initial context.

Contributions

This thesis is based on the following journal papers, the correspondance between papers and thesis sections is displayed in table 2.

Chapter 2	"Binary Classification based Monte Carlo sampling"	Section 2.1
Chapter 3	"Generative vs. Discriminative Bayesian Posterior learning"	-
Chapter 4	"Discretely Indexed Flows"	Section 4.3

Table 2: Correspondance between papers and sections

- "Binary Classification Based Monte Carlo Simulation", **Elouan ARGOUARC'H**, François DESBOUVRIES; published in IEEE Signal Processing Letters, vol. 31, pp. 1449-1453, 2024, doi: 10.1109/LSP.2024.3396403.

  https://ieeexplore.ieee.org/document/10517652;

- "Generative vs. Discriminative Bayesian Posterior learning", **Elouan ARGOUARC'H**, François DESBOUVRIES, Eric BARAT, Eiji KAWASAKI; submitted to *Bayesian Analysis*. Available on ArXiv:stat/2406.09172. - https://arxiv.org/abs/2406.09172;

- "Discretely Indexed Flows", **Elouan ARGOUARC'H**, François DESBOUVRIES, Eric BARAT, Eiji KAWASAKI, Thomas DAUTREMER; working paper. Available on ArXiv:stat/2204.01361. - https://arxiv.org/abs/2204.01361.

In addition to these three articles, preliminary versions of the work were also published in a national signal and image processing conference. They are related to this thesis, respectively parts of chapter 4 and chapter 3.

- "Flots stochastiques discrets" **Elouan ARGOUARC'H**, François DESBOUVRIES, Eric BARAT, Eiji KAWASAKI, Thomas DAUTREMER; Actes du 28ème colloque GRETSI, Nancy, France, September 2022.

  https://www.gretsi.fr/data/colloque/pdf/2022_argouarch1049.pdf;

- “Apprentissage Bayésien Semi-supervisé par modélisation générative”, **Elouan ARGOUARC’H**, François DESBOUVRIES, Eric BARAT, Eiji KAWASAKI, Thomas DAUTREMER; Actes du 29ème colloque GRETSI,Grenoble, France, September 2023.

  https://www.gretsi.fr/data/colloque/pdf/2023_argouarch1365.pdf.

Contents

Chapter 1

From Bayesian Inference to Posterior Statistical Learning

In this introductory chapter, we propose a guided tour from classical Bayesian inference to posterior statistical learning. We indeed aim to explain how considering increasingly intricate models can render the usual likelihood-based posterior inference methods unfeasible, and how an appropriate use of learning techniques enables us to bypass such shortcomings. This establishes the context of this thesis, and enables us to raise a number of questions which will be addressed in the following chapters.

Consider a scientific problem where we dispose of recorded observations, and the goal is to study the underlying phenomenon by interpreting some underlying causes or properties, or by predicting future outcomes of the same phenomenon. This general problem can be tackled with the framework of Bayesian posterior inference (114)(4) which we now describe. We represent the unknown phenomenon of interest using a probabilistic model based on two random variables (RVs). More precisely, we associate the observed RV Y to a hidden interpretable RV of interest X, which we assume (i) is distributed according to $\mathcal{P}_X$ and (ii) is related to Y via an *observation model* $\mathcal{P}_{Y|X}$. Observing the value of Y, say y, indeed carries information about X, which is encapsulated in the distribution $\mathcal{P}_{X|Y=y}$ and with a probability density function[1] (PDF) given by the Bayes formula (3):

$$p_{X|Y=y}(x) = \frac{p_X(x)p_{Y|X=x}(y)}{\int p_X(x)p_{Y|X=x}(y)\mathrm{d}x}. \tag{1.1}$$

$\mathcal{P}_{X|Y=y}$ is the *a posteriori* (or posterior) distribution and is the result of Bayesian updating: the *a priori* (or prior) knowledge about X described by $\mathcal{P}_X$ is updated with the observation that Y takes the value y via the *likelihood*[2] $p_{Y|X=x}(y)$.

If the model is accurate enough, so that it reasonably describes the natural phenomenon of interest, then the posterior distribution can indeed provide valuable insights about the probable values of X given $Y = y$, and thus to interpret the probable causes

[1] Throughout this thesis, the PDF should be understood with respect to the appropriate measure, depending on the nature (continuous, discrete, or mixed continuous-discrete) of the underlying variables.

[2] In the formula for the posterior PDF, the variable of interest is x, and so even though the likelihood is the PDF of a distribution over Y evaluated at the observed value y, it is indeed to be understood as a function of x.

of the underlying physical phenomenon. We now illustrate this point with a personally made-up (and possibly unrealistic) example.

Example: *Say we observe a tree at some geographical location and we wish to understand how such a big (or tiny) tree came to grow in this specific area. The observed RV is* $Y = \{size, location\}$ *and takes value* $y = \{3m \times 1.5m, Palaiseau - France\}$. *For this problem, a relevant cause might be the specie and age of the tree. So we chose to consider the RV* $X = \{specie, age\}$ *where specie can be a Categorical variable which takes values among the list of known tree species, and age can be a continuous positive variable. Consider the distribution* $\mathcal{P}_X$ *which can describe prior information such as: "Oak is more common than Birch", "an Eucalyptus usually lives between 150 and 700 years", "a 200 year old tree is more likely to be a marple tree than a cherry tree", and "the average lifespan of a tree is around 350 years". On the other hand, we assume an observation model* $\mathcal{P}_{Y|X}$ *which indeed relates* Y *to* X *with information such as: "Spruce thrives more in northern areas", "Sequoias are more likely to reach heights beyond 80 meters than pines", and "the size of a tree increases with age". Then, by examining the posterior probability distribution, we might, for example, conclude that, under these probabilistic assumptions, it is more likely that the observed tree is a mature Hazelnut tree rather than a young Linden tree.*

1.1 Monte Carlo Integration

In turn, studying the properties of the posterior distribution can, most of the time, be expressed as a problem of computing (or approximating - and possibly minimizing) an expectation of the form:

$$\mathbb{E}_{X\sim\mathcal{P}_{X|Y=y}}[f(X)]. \tag{1.2}$$

where f is a measurable function. Computing this integral enables us to extract meaningful information about the RV X, once is observed that $Y = y$. For instance, let A be a set, then with $f(x) = \mathbb{1}_A(x)$, this expectation becomes $\Pr(X \in A|Y = y)$. This procedure therefore enables us to discover the probable regions ($\sim$ values) for the RV X, given the observation of the random process of interest. Another example is the case where $f(x) = x$, then this expectation becomes $\mathbb{E}[X|Y = y]$ and corresponds to the average value of X given $Y = y$. Finally, the posterior distribution enables to obtain pointwise Bayes predictors by minimizing the expectation of a well-designed loss function l (80): $x^* = \arg\min_x \mathbb{E}_{X\sim\mathcal{P}_{X|Y=y}}[l(X, x)]$; which indeed corresponds to minimizing an expectation of the same form as in (1.2).

Hence, we consider the goal of computing an expectation computed with respect to the posterior, as in (1.2), with a given function f. This expectation is written as, using the *Law of the unconscious statistician*:

$$\mathbb{E}_{X\sim\mathcal{P}_{X|Y=y}}[f(X)] = \int f(x)p_{X|Y=y}(x)\mathrm{d}x. \tag{1.3}$$

Unfortunately, for an arbitrary function f, this integral does not admit an analytical closed form expression such that no feasible computation can yield a numerical value. So

it cannot be computed exactly, and this shortcoming instead calls for an approximation of the integral.

The problem of approximating an integral does not only occur when tackling Bayesian posterior inference, and is a more general problem which is widely studied; see e.g. (78)(51). The possible methods to provide such an approximation include the Riemann summation methods as well as many other schemes which consist in approximating the integral as a weighted sum of the integrand evaluated at a finite set of points. We refer to such methods as standard numerical integration methods as opposed to Monte Carlo integration, and we refer to the book (25) for a complete overview of these methods. However, these standard numerical integration methods are not best suited for approximating (1.2) as, firstly, they fail to provide an accurate approximation when the integral is computed on (i) an infinite and/or (ii) a multidimensional set. Indeed, to reach a given precision of the estimate, the number of required integrand evaluations grows exponentially with the dimension. Secondly, and perhaps more critically, the standard numerical integration methods cannot even be applied in this context since the integrand function cannot be exactly evaluated, and we now explain this precise point.

In equation (1.3), the integrand is a product of two functions: the first factor is the measurable function $f(x)$, and the second is the posterior PDF $p_{X|Y=y}(x)$. It turns out that standard numerical methods cannot be applied in this setting since this integrand can only be evaluated up to a constant as the second factor, which is given by expression (3.1), cannot be evaluated. Indeed, on the one hand, if the likelihood and prior are conjugate (41) (see also (17) and (23) for relevant references), then the posterior distribution can be expressed as some known distribution which belongs to the same family as the prior, and as such its PDF can be expressed and computed exactly. However, conjugacy usually occurs when a practitioner specifically selects the likelihood and prior to be conjugate for one another for computational convenience; this case rarely corresponds to a practical application where one or both can be arbitrary probability distributions. Indeed, for an arbitrary choice of $\mathcal{P}_{Y|X}$, (i) a known conjugate prior is not guaranteed to exist (as is the case with a logistic model for instance - though it is always the case for exponential family distributions (30)) and (ii) the distribution $\mathcal{P}_X$ does not necessarily correspond to one such conjugate prior. As a consequence, in general, only the numerator of the posterior PDF $p_X(x)p_{Y|X}(x|y)$ can be evaluated, while its constant denominator $\int p_X(x)p_{Y|X}(y|x)\mathrm{d}x$ is itself an intractable integral. So, we can write the posterior PDF as:

$$p_{X|Y=y}(x) = \frac{\widetilde{p}_{X|Y=y}(x)}{C}, \tag{1.4}$$

where only $\widetilde{p}_{X|Y=y}(x)$ can be evaluated and $C = \int \widetilde{p}_{X|Y=y}(x)\mathrm{d}x$ is the intractable normalizing constant.

Monte Carlo integration refers to a set of numerical methods for approximating integrals which allows us to bypass the shortcoming that the integrand is intractable in general. Indeed, it provides an approximation of (1.2) that is not based on a linear combination of the integrand but which is instead based on random number generation.

The principle of this approximation is as follows:

$$\mathbb{E}_{\mathcal{P}_{X|Y=y}}[f(X)] \approx \frac{1}{N}\sum_{i=1}^{N} f(X_i);\ \text{where } X_1, ..., X_N \overset{iid}{\sim} \mathcal{P}_{X|Y=y}. \tag{1.5}$$

The weak (resp. strong) *law of large numbers* states that the right-hand side of (1.5) converges in probability (resp. almost surely) to the true value of the expectation. Moreover, the central limit theorem implies that, with a probability close to 0.95:

$$|\mathbb{E}_{\mathcal{P}_{X|Y=y}}[f(X)] - \frac{1}{N}\sum_{i=1}^{N} f(X_i)| \leq \frac{1.96\sigma_X}{\sqrt{N}} \tag{1.6}$$

where $\sigma_X^2 = \mathbb{V}\text{ar}[X_i]$. This means that the Monte Carlo integration principle yields an approximation of the expectation with absolute error bounded, with high probability, by $\mathcal{O}(\frac{1}{\sqrt{N}})$ and so the convergence rate of this method is $\sqrt{N}$ regardless of the dimension.

1.2 Sampling from the posterior

With Monte Carlo integration, the problem of approximating an integral has thus become a problem of sampling, and our ability to obtain an approximation of (1.2) is directly related to our ability to draw independent samples from the posterior distribution.

As we have mentioned before, the posterior distribution can be expressed as some known probability distribution only in specific cases, such as when the prior and the likelihood are conjugate, in which case it belongs to the same family as the prior distribution. If this family of distributions benefits from a specific sampling procedure (52)(28), then the posterior distribution inherits from this computational advantage. In the general case, the posterior distribution is only described via its PDF (up to an unknown constant) given by the Bayes formula (3.1), which raises the general question of sampling from this unnormalized PDFs.

We now briefly review the historical methods for Monte Carlo sampling which enable us to approximate an expectation (1.2) (in this section, and unless stated otherwise, the distribution of interest is not necessarily a posterior of the form $\mathcal{P}_{X|Y}$ but is more generally a distribution over X which we denote by $\mathcal{P}$).

1.2.1 Accept-Reject Sampling

Accept-reject sampling, or rejection-sampling, is a technique to sample from a distribution of interest, which is known via its PDF. It can be traced back to paper (128) and has rapidly become a cornerstone of random sampling (27) (see e.g.(92) with a chapter dedicated to a recent, complete and thorough review of the technique) and has led to the development of several ensuing methods such as, most notably, (70). The seminal technique of rejection sampling is simple and elegant (43): its principle is to propose a sample from a suitable proposition distribution and draw a binary RV which indicates

whether to accept the proposed sample or to reject it (hence the name *Accept-Reject* sampling). More precisely, consider an instrumental distribution $\mathcal{Q}$ which dominates the distribution of interest $\mathcal{P}$: for any set A, $\mathcal{P}(A) > 0$ implies $\mathcal{Q}(A) > 0$. As a consequence, if $p(x) > 0$ then $q(x) > 0$ which subsequently implies that $c^* \triangleq \sup_{x \in \mathbb{R}} \frac{p(x)}{q(x)} \geq 1$. Consider (i) $c \geq c^*$, (ii) X a random proposed sample distributed according to the easy-to-sample proposal $\mathcal{Q}$ and (iii) k a Bernoulli RV with probability $\frac{p(X)}{cq(X)}$. The score point of rejection sampling is that, given $k = 1$, X is exactly distributed according to $\mathcal{P}$ which we can easily see as its PDF reads:

$$\frac{q(x)\Pr(k=1|x)}{\Pr(k=1)} = \frac{q(x)\frac{p(x)}{cq(x)}}{\int q(x)\frac{p(x)}{cq(x)}\mathrm{d}x} = p(x). \tag{1.7}$$

The rejection sampling algorithm can therefore be applied when one can compute the PDF of the target distribution as well as that of the proposal instrumental distribution. However, this technique can easily be adapted to the case where p and/or q are each known up to a constant (which is not necessarily common), as is explained in (100). The acceptance probability is $\Pr(k = 1) = \frac{1}{c}$. So the lower the value of c, the more likely it is that a proposed sample will get accepted: the efficiency of the rejection sampling approach is dictated by (i) the ability of a practitioner to elicit a suitable proposal distribution which is close to the target distribution, yielding a small value of c^* and (ii) the ability to find a low value of $c \geq c^*$. With that regard, one can use different proposal distributions (perhaps from the same parameterized family) during the sampling procedure and, with some precautions, still obtain independent samples as in (12). In the special case of a log-concave target PDF, it is possible to use a piecewise linear proposal distribution, which we can refine during the accept-reject sampling procedure, yielding an adaptative rejection sampling procedure (54).

1.2.2 Markov chain Monte Carlo

Markov chain Monte Carlo (see (7) for a review) is a set of techniques for sampling from a target distribution by simulating a Markov chain which admits that distribution as its limiting invariant distribution. MCMC techniques therefore take interest in designing a Markov transition kernel $\mathcal{M}$ which is easy-to-sample from and which leaves the target distribution $\mathcal{P}$ invariant:

$$\mathcal{P}(\mathrm{d}x) = \int \mathcal{M}(x', \mathrm{d}x)\mathcal{P}(\mathrm{d}x'). \tag{1.8}$$

Simulating the Markov chain with transition kernel $\mathcal{M}$ therefore provides correlated samples that are asymptotically distributed according to $\mathcal{P}$. A sufficient condition for $\mathcal{M}$ to leave $\mathcal{P}$ invariant is that $\mathcal{M}$ is *reversible* with respect to $\mathcal{P}$ (126):

$$\mathcal{M}(x', \mathrm{d}x)\mathcal{P}(\mathrm{d}x') = \mathcal{M}(x, \mathrm{d}x')\mathcal{P}(\mathrm{d}x). \tag{1.9}$$

Indeed, one can easily check that this condition, which is also referred to as *detailed balance* or *time reversibility*, satisfies (1.8). For example, if $\mathcal{P}$ is a multivariate distribution, then sampling from the probability distribution of one variable conditionally on

the values of the others is indeed a reversible Markov transition kernel. This principle yields the Gibbs Sampling algorithm (50)(48), which consists in, sequentially or in a random order, sampling some or all the conditional distributions. Note finally that, even though Gibbs sampling is a technique which is suited for sampling from multivariate distributions, its principle can be of interest for sampling from univariate distributions: one can utilize the Gibbs sampling procedure to an arbitrary joint distribution which admits $\mathcal{P}$ as one of its marginals. A most notable application of this principle is that of Slice-Sampling (24)(99), which can easily be understood as Gibbs sampling applied to an augmented (with a uniform RV) distribution, whose PDF reads $p(x,u) = p(x)\frac{\mathbb{1}_{[0,p(x)]}(u)}{p(x)}$.

The Metropolis-Hastings (MH) algorithm (93)(59), perhaps the most famous MCMC algorithm, is another very convenient approach to construct a transition kernel which satisfies the detailed balance property. The idea of the MH algorithm is to propose a candidate x^* according to some Markov kernel $\mathcal{Q}$ (not necessarily $\mathcal{P}$-invariant or $\mathcal{P}$-reversible) and either accept the proposed point (in which case it becomes the next state of the Markov chain) or reject it (in which case the chain remains at the current point x_t) according to a Bernoulli distribution with acceptance probability $\alpha_{MH}(x^*, x_t) = \min(1, \frac{p(x^*)q(x_t|x^*)}{p(x_t)q(x^*|x_t)})$. It happens that the resulting two-step transition indeed forms a Markov transition kernel which is reversible with respect to $\mathcal{P}$ (see (18) for a comprehensive review of the MH acceptance). The interest of this MH scheme is that one can use any transition kernel. For instance, kernels can be informed by the geometry of $\mathcal{P}$ using the gradient of the log-PDF to drive the chain towards regions of high mass, which is the idea of Metropolis-Adjusted Langevin (57)(116)(115) and Hamiltonian Monte Carlo (35)(98)(63). Finally, observe that the MH scheme is not the only way to construct a reversible Markov kernel, as alternative acceptance probabilities also ensure reversibility (2)(83)(79), such as Barker's acceptance.

1.2.3 Importance Sampling

Importance sampling is a Monte Carlo integration technique which can be used to approximate integrals such as (1.2) and which can be traced back to the late 1940's (68),(55) and the early 1950's (67),(91),(58). This technique is based on the simple rewriting of the expectation of interest as:

$$\mathbb{E}_{X\sim\mathcal{P}}\big[f(X)\big] = \mathbb{E}_{Z\sim\mathcal{Q}}\left[f(Z)\frac{p(Z)}{q(Z)}\right]; \tag{1.10}$$

where $\mathcal{Q}$ is an instrumental distribution called the importance distribution. Since we have rewritten the expectation computed with respect to $\mathcal{P}$ as an expectation computed with respect to $\mathcal{Q}$, one can obtain a Monte Carlo approximation of the expectation by sampling from the importance distribution as:

$$\mathbb{E}_{X\sim\mathcal{P}}\big[f(X)\big] \approx \frac{1}{N}\sum_{i=1}^{N} f(Z_i)\frac{p(Z_i)}{q(Z_i)}, \text{where } Z_1, ..., Z_N \overset{iid}{\sim} \mathcal{Q}. \tag{1.11}$$

This technique is, of course, of particular interest when we cannot sample from $\mathcal{P}$ directly, in which case we cannot even construct the vanilla Monte Carlo estimate

$\frac{1}{N}\sum_{i=1}^{N} f(X^{(i)})$, where $X^{(1)}, ..., X^{(N)} \overset{iid}{\sim} \mathcal{P}$. When we are interested in finding an importance distribution, or even reusing the same set of samples from that importance distribution, in order to approximate expectations (1.2) for a broad spectrum of functions f, then it is of interest to consider an importance distribution which is as close as possible to $\mathcal{P}$, see (11)(32). But we now see that this technique can also be of interest in the case where we are able to obtain samples from $\mathcal{P}$ but the corresponding vanilla estimator is a poor approximation for a fixed number of samples and for a given function f. Indeed, importance sampling can be understood as a variance reduction technique since an appropriate choice of importance distribution enables us to obtain a Monte Carlo estimate of low variance. To understand this, let us examine the variance of the importance sampling estimate (1.11) with importance distribution $\mathcal{Q}$ for a real-valued function f which reads:

$$\frac{1}{N}\left(\mathbb{E}_{Z\sim\mathcal{Q}}\left[\left(f(Z)\frac{p(Z)}{q(Z)}\right)^2\right] - \mathbb{E}_{X\sim\mathcal{P}}\big[f(X)\big]^2\right). \tag{1.12}$$

We see that an appropriate choice of importance distribution $\mathcal{Q}^*$ with PDF $q^*(x) = \frac{|f(x)|p(x)}{\int |f(x)|p(x)\mathrm{d}x}$ yields an estimate of minimal variance (which even reaches 0 when $f > 0$). Even if $\mathcal{Q}^*$ is not an easy-to-sample-from distribution which can be used in practice as importance distribution, this tells us that the regions where it is most important to obtain samples from (hence the term *importance* distribution) are not those where only $p(x)$ is large, but rather those where $|f(x)|p(x)$ is large.

When the probability distribution $\mathcal{P}$ and/or the importance distribution $\mathcal{Q}$ have PDFs which are known up to constants, say $\widetilde{p}$ and $\widetilde{q}$ (as is the case in (1.2) where the expectation is computed with respect to (3.1)), one can resort to self-normalized importance sampling estimation (53)(8). This estimation method is based on the rewriting of the expectation of interest as:

$$\mathbb{E}_{X\sim\mathcal{P}}\big[f(X)\big] = \frac{\mathbb{E}_{Z\sim Q}\left[f(Z)\frac{\widetilde{p}(Z)}{\widetilde{q}(Z)}\right]}{\mathbb{E}_{Z\sim Q}\left[\frac{\widetilde{p}(Z)}{\widetilde{q}(Z)}\right]}; \tag{1.13}$$

We can build an estimate of this expectation by estimating both the numerator and denominator with the importance sampling principle with the same set of samples $Z_1, ..., Z_M \overset{iid}{\sim} \mathcal{Q}$ (i.e. with the same samples from the same importance distribution, though it can also be done using different importance distributions for the numerator and the denominator, see (77) for an application of this idea). This yields:

$$\mathbb{E}_{X\sim\mathcal{P}}\big[f(X)\big] \approx \sum_{i=1}^{N} \frac{\widetilde{\omega}_i}{\sum_{j=1}^{N}\widetilde{\omega}_i} f(Z_i), \text{where } \widetilde{\omega}_i = \frac{\widetilde{p}(Z_i)}{\widetilde{q}(Z_i)}; \tag{1.14}$$

which, unlike the unnormalized importance sampling estimate (1.11), is biased but is nonetheless asymptotically unbiased. Moreover, an inspection of the variance of this estimate yields an optimal importance distribution, which PDF reads $q^*(x) =$

$\frac{|f(x)-\mathbb{E}_{X\sim\mathcal{P}}\left[f(X)\right]|p(x)}{\int|f(x)-\mathbb{E}_{X\sim\mathcal{P}}\left[f(X)\right]|p(x)\mathrm{d}x}$. Unlike in the first case, where the optimal importance distribution was only inconvenient to use since it could not be sampled from easily, the self-normalized optimal importance distribution cannot be used in practice since it involves the unknown value of the expectation that we are trying to approximate in the first place. The properties of this estimate, both asymptotic (61)(113)(82) and non-asymptotic (1)(16), have been thoroughly studied and are well understood.

Importance sampling can also be understood as an implicit sampling mechanism. Indeed, the importance sampling estimate (1.14) can be understood as an exact expectation computed with respect to a discrete measure in the form of a weighted sum of Dirac-δ functions:

$$\sum_{i=1}^{N}\omega_i\delta_{Z_i}(\mathrm{d}x), \text{where } \omega_i=\frac{\widetilde{\omega}_i}{\sum_{j=1}^{N}\widetilde{\omega}_i} \text{ and } Z_1,...,Z_N\overset{iid}{\sim}\mathcal{Q}; \tag{1.15}$$

from which one can easily obtain samples via a resampling procedure:

$$Z_j \text{ where } j\sim\text{Categorical}(\omega_1,...,\omega_N). \tag{1.16}$$

This is the principle of Rubin's Sampling-Importance Resampling (117)(122). Repeating this sampling procedure produces samples which are correlated and approximately distributed according to $\mathcal{P}$, and of course, since we are resampling (with replacement) amongst a finite set of samples, we can obtain several replicas of the same sample from the importance distribution. The score point of sampling-importance-resampling is that the produced samples become iid samples from $\mathcal{P}$ as the number of proposed samples N increases to infinity (9, chapter 9).

1.2.4 Variational Inference

Variational Inference (VI) (66) (see (5) for a gentle introduction) methods consist in approximating $\mathcal{P}$ with an instrumental distribution $\mathcal{Q}^*$ obtained via minimizing a discrepancy measure $\mathcal{D}^{\text{VI}}$ over a family of distributions $\mathcal{F}$:

$$\mathcal{Q}^*=\arg\min_{\mathcal{Q}\in\mathcal{F}}\mathrm{D}^{\text{VI}}(\mathcal{P},\mathcal{Q}). \tag{1.17}$$

If $\mathcal{F}$ represents a set comprised of easy-to-sample-from distributions, then $\mathcal{Q}^*$ can easily provide samples that are approximately distributed under $\mathcal{P}$. Therefore, VI turns a problem of sampling from a distribution known via its PDF into a problem of discrepancy minimization, usually over a parameterized family of distributions (hence the term *variational* inference). When applied to a posterior distribution, VI enables to perform approximate Bayesian posterior inference (hence the name variational *inference*) (44), which finds applications, for instance, in the learning of implicit generative models (see e.g. (72)). In the more general unconditional setting, this method is, by abuse of language, still referred to as VI.

Once the variational distribution is obtained via optimization, $\mathcal{Q}^*$ can be used as an approximation of the target distribution $\mathcal{P}$; alternatively, $\mathcal{Q}^*$ can be used as an optimized

instrumental distribution in a classical Monte Carlo sampling setting, in order to compensate for the discrepancy. For instance, (10) proposes an Expectation-Maximization (26) algorithm with closed-form updates to sequentially adjust the importance distribution. Recent advances in gradient-based optimization (71), the development of flexible (neural network) based generative models such as Normalizing Flows (111)(102), together with, if necessary, a reparameterization of gradients (40)(132), as an alternative to the previous log-trick (130), have enabled practitioners to construct and sample from variational approximations of complex distributions. Current challenges in VI (29) include (i) eliciting an appropriate discrepancy measure D^{VI} (81)(129)(47), (ii) a suitable generative modeling technique to constitute $\mathcal{F}$ (73)(13) which (iii) can be sampled from via a reparametrizable scheme (125)(97)(135)(108)(22).

1.3 From Bayesian Inference to Statistical Learning

Let us now go back to the general principles of Bayesian posterior inference. The score point is to provide a relevant modeling of a natural phenomenon (be it physical, biological, economical, environmental, behavioral...), say $\mathcal{P}_0$. Due to its complexity, scientists often resort to adopting a hypothesis of causality: there are *causes* to the observed phenomenon. This leads the practitioner to introduce some variable X which aims to describe such causes, and his/her goal becomes the study of the probable causes, given observations y, via an appropriate conditional distribution $\mathcal{P}_{X|Y=y}$.

The Bayesian posterior framework indeed enables us to construct such a distribution, by building a probabilistic model with a two-step procedure. We first seek exogenous prior information about the causes, which we transcribe into a prior distribution associated to RV X (112). We then specify an observation model $\mathcal{P}_{Y|X}$: a conditional distribution which represents the assumed relationship from causes to observations. In turn, once we record observations $Y = y$ from $\mathcal{P}_0$, we retrieve the probable values x of X, or equivalently, the probable models $\mathcal{P}_{Y|X=x}$ for $\mathcal{P}_0$. Finally this methodology enables either (i) to understand (or at least to interpret) the probable causes or properties of $\mathcal{P}_0$ under the scope of the considered model with an interpretable X; or (ii) to predict future outcomes, say Y', of the same unknown phenomenon by examining the predictive distribution with PDF $p_{Y'|Y=y}(y') = \int p_{Y'|X=x}(y')p_{X|Y=y}(x)\mathrm{d}x$.

As we have mentioned in the previous section, the ability to extract meaningful information via Bayesian inference depends on computational considerations. However, the relevance of the underlying inference problem is also related to the considered probabilistic modeling and we now discuss this point. Assuming the observation model (alongside the prior distribution) amounts to defining a set of probability distributions $\mathcal{S} = \left\{\mathcal{P}_{Y|X}, X \sim \mathcal{P}_X\right\}$ indexed by the values of X. Two different cases then arise: (i) if $\mathcal{P}_0 \in \mathcal{S}$, or equivalently if there exists a value x_0 such that $\mathcal{P}_{Y|X=x_0}$ corresponds to the distribution $\mathcal{P}_0$, we say that the observation model is well specified; (ii) otherwise, if such a value of x_0 does not exist and $\mathcal{P}_0 \notin \mathcal{S}$, we say that it is misspecified. Well-, or mis-, specification is a result of the choice of the observation model and is often determined by our exogenous knowledge of the underlying random process. For instance, if we know for certain that $\mathcal{P}_0$ is a Gaussian distribution with unknown mean and vari-

ance parameters, then, on the one hand, modeling $\mathcal{P}_0$ with a Gaussian distribution $\mathcal{P}_{Y|X} = \mathcal{N}(\mu, \sigma^2)$ where $X = \{\mu, \sigma^2\}$ is a well-specified setting, while modeling $\mathcal{P}_0$ with a Laplace distribution $\mathcal{P}_{Y|X} = \mathcal{L}(\mu, b)$ where $X = \{\mu, b\}$ is a misspecificed setting.

Under an assumption of identifiability ($x_1 \neq x_2$ implies $\mathcal{P}_{Y|X=x_1} \neq \mathcal{P}_{Y|X=x_2}$) and mild regularity conditions, as the number of independent observations from $\mathcal{P}_0$ increases, the posterior distribution converges to a point mass Dirac distribution (see (49, chapter 4)) at:

$$x^* = \arg\min_x \mathrm{D}_{\mathrm{KL}}(\mathcal{P}_0 || \mathcal{P}_{Y|X=x}); \tag{1.18}$$

where $\mathrm{D}_{\mathrm{KL}}(\mathcal{P}||\mathcal{Q}) = \mathbb{E}_{X\sim\mathcal{P}}\big[\log(p(X)) - \log(q(X))\big]$ is the Kullback-Leibler divergence (76)(75). Moreover, the posterior is also asymptotically normal, centered at x^* and with covariance matrix $(n\mathcal{I}(x))^{-1}|_{x=x^*}$ where $\mathcal{I}$ is the Fisher Information matrix (42) about model $\mathcal{P}_{Y|X}$. Of course, in the well-specified case, x^* corresponds to x_0, which is the value that frequentists consider as the true unknown value of interest. So, as the number of observations increases, Bayesian inference (credible intervals) is consistent with the frequentist approach (confidence intervals) in the well-specified case. However, in real-world applications, one might expect most, if not all, Bayesian problems to be misspecified. Indeed, since $\mathcal{P}_0$ is only available via its recorded observations and is otherwise never precisely known (otherwise we would not need to study this process via Bayesian inference in the first place), we can never ensure that we select a model such that $\mathcal{S}$ includes $\mathcal{P}_0$, nor can we confirm that $\mathcal{P}_0$ indeed belongs to a given $\mathcal{S}$. As George Box states it in (6):

> "All models are wrong."

Nonetheless, even though it is inevitably stained with some inaccuracy, making a relevant assumption about the relationship between Y, and a considered hidden X using an appropriate observation model $\mathcal{P}_{Y|X}$, remains a crucial aspect of Bayesian posterior inference. Through studies and experiments, scientists and practitioners can improve their understanding of the underlying phenomena and physical mechanisms, and can therefore in turn enrich the physical models use for inference and thus reduce the effect of misspecification. However, precise representations of intricate phenomena is often reliant on evermore complex models which can lead to further limitations in the context of Bayesian posterior inference. Indeed, in many scientific applications (famous examples include population fluctuation (84)(127) and compartmental modeling in epidemiology (69)), we may have strong arguments in favor of a specific observation model $\mathcal{P}_{Y|X}$ for its otherwise scientific relevance, but in such cases the risk of more and more precise modeling is that the corresponding distribution does not necessarily provide a tractable PDF any longer. This setting is referred to as *likelihood-free* and this is the context considered throughout the rest of this thesis.

In the rest of this section, we first explicit the case where the observation model is defined as a generative model ($\sim$ a simulator - see section 1.3.1), and present the gold standard methods of ABC in this context (see section 1.3.2). We then explicit the situations where (i) the observation model is with an intractable PDF and moreover (ii) ABC is unfeasible; which finally leads us to presenting the posterior learning problem (see section 1.3.3).

1.3.1 Observation model as a simulator with intractable PDF

In Bayesian inference, as soon as we depart from simple observation models $\mathcal{P}_{Y|X}$, we often face the major drawback that the corresponding PDF $p_{y|x}(y)$ is intractable, costly to evaluate or accurately approximate, or too noisy to efficiently work with (131)(38)(104).

Before adressing the general setting of this thesis where the observing model is only available via a recorded dataset (see section 1.3.3), let us first focus on the specific setting where the observation model is a *generating process*. This means that it is defined as, and perhaps only available via, a sampling procedure which takes as input x and outputs a value y via a succession of operations: deterministic or stochastic mathematical computations, numerical or computer-based operations, practical or thought experiments conducted by an operator, or even a recorded random process from nature. In this case, more often than not, the observation model has an intractable PDF. We sometimes describe as *implicit* such distributions, which indeed can be sampled from, but the corresponding PDF is not available in closed-form.

A most notable example is when the observation model is a simulator which involves a call to a (pseudo-) random number generator. Indeed, if a computer simulator involves drawing (one or several) latent RVs, then its PDF is computed by marginalizing out the latent variables $p_{Y|X=x}(y) = \int p_{Y,Z|X=x}(y,z)\mathrm{d}z$. In general, for complex simulation procedures, this expression cannot be expressed in closed form nor can it be estimated accurately either. In particular, if, for example, the simulation process involves a non-invertible function, say $y = h(z)$, then the likelihood function $p_{Y|X=x}(y) = \int \delta_{h(z)}(y)p_{Z|X=x}(z)\mathrm{d}z$ cannot be evaluated as the integral cannot be computed using the change of variables technique.

The case where this PDF is intractable also occurs when $\mathcal{P}_{Y|X}$ is a real-world experiment: a measurement from some intrinsic random process from nature, be it a practical experiment or a thought experiment, where X represents the experimental design and Y the measured or recorded output. Generally speaking, such a real-world process can usually not be described, or at least not accurately enough, in terms of a tractable probability distribution because of our ignorance or limited understanding of the underlying generation mechanism.

This difficulty heavily hinders our ability to perform efficient Bayesian posterior inference, as in this case, the posterior probability distribution has a PDF $p_{X|Y=y}(x)$ that cannot be evaluated, not even up to its normalizing constant. Indeed, as we have mentioned in the corresponding section, Bayesian inference relies on evaluation of the posterior PDF which itself involves evaluating the likelihood function via equation (3.1). This setting is often considered in the literature and is referred to as a Likelihood-free setting (31)(118)(34)(38)(87) which naturally arises in many different fields such as (56) in econometrics, (124)(90) in molecular genetics, (110) in epidemiology, (106) protein evolution.

1.3.2 Approximate Bayesian Computation methods

In the setting where the observation model is available via its sampling mechanism but has unavailable PDF (as mentionned in the previous section), historical methods include Approximate Bayesian Computation (ABC) techniques (see article (88) and book (120) for a review of challenges and advances), and, for context, we briefly present the seminal approach of rejection-ABC in this section.

ABC methods have came a long way and can be traced back to the paper (118) in which what would later become the seminal rejection-ABC method (124) is rather used as an intuitive explanation on how the prior distribution and the likelihood (the PDF associated with the observation model) interact to produce the posterior distribution. Consider the following algorithm 1.

Algorithm 1 Rejection ABC

Require: observed y, $\epsilon > 0$, kernel $K_{\_}$
  while No value is accepted **do**
    propose $x' \sim \mathcal{P}_X$
    simulate $y' \sim \mathcal{P}_{Y|X=x'}$
    if $u \sim \mathcal{U}_{[0,1]} \leq K_\epsilon(y', y)$ **then**
      Accept x'
    end if
  end while

In this algorithm $K_\epsilon(.,.)$ is a kernel function with bandwidth ϵ and measures the discrepancy between y and y'. Examples of such kernel functions include the Gaussian kernel where $K_\epsilon(y, y') = \mathcal{N}(y - y'; 0, \epsilon \mathrm{I})$. Therefore, the closer y' is to y, the higher the probability that x' is accepted. The accepted samples are drawn from a distribution $\mathcal{P}_\epsilon^{ABC}$ with PDF :

$$p_\epsilon^{ABC}(x') = \int K_\epsilon(y', y) p_{Y|X=x'}(y') p_X(x') \mathrm{d}y'; \tag{1.19}$$

and the score point of ABC is that, if $\lim_{\epsilon \to 0} K_\epsilon(y', y) = \delta_y(y')$ then:

$$\lim_{\epsilon \to 0} \mathcal{P}_\epsilon^{ABC} = \mathcal{P}_{X|Y=y}. \tag{1.20}$$

So for small values of ϵ, the rejection ABC indeed provides samples which are approximately distributed under the posterior distribution $\mathcal{P}_{X|Y=y}$ without requiring to evaluate the intractable PDF of the observation model. The pitfall of this ABC algorithm is that, the smaller the value of ϵ, the smaller the probability that a proposed y' is close to y, within the probable range controlled by ϵ, and thus that x' is accepted. This effect is amplified when the dimension of the observation y increases (the intrinsic dimension and the number of observations). Several refinements were proposed to increase the acceptance rate of this ABC scheme which include most notably: (i) the use of summary statistics (39), facilitating dealing with high dimensional data more conveniently, and (ii) the

coupling of ABC with more refined Monte Carlo sampling schemes such as MCMC (89), Gibbs sampling (20), SMC (121), resulting in an improved exploration of the X-space. While concurrent approaches based on statistical learning of the unavailable posterior enabled by recent advances in machine-learning are becoming increasingly popular, as discussed in the following section, ABC methods have been thoroughly studied and applied successfully in many scientific problems; they are still considered by practitioners as a gold standard in many practical likelihood-free settings.

1.3.3 Statistical learning of a model for the posterior

We now describe two settings where ABC is no longer a viable option for sampling from the unknown posterior distribution in a likelihood-free setting.

On the one hand, it is possible that the observation model used to be available via its generation mechanism, but is currently no longer accessible. This happens for example when the experimental setting has changed and that the conditions which previously enabled conducting the experiment (simulation) are no longer met. ABC methods cannot be applied since we can no longer simulate y' from the observation model for a given proposed input x'. However, it is possible that previous recording indeed provided $\mathcal{D} = \left\{(x_i, y_i), y_i \sim \mathcal{P}_{Y|X=x_i}\right\}$. This situation corresponds to the most common setting which we encounter in classical classification and regression tasks where we dispose of a finite dataset which was generated beforehand via an unknown data generation process.

On the other hand, it is possible that the observation model remains available via its sampling mechanism, so that obtaining a simulated y for a given x is feasible, but the process is resource-intensive. In this case, ABC methods might not be a viable approach either since it can be prohibitively costly to compensate for poor exploration of the x-space (which yields low acceptance rates in rejection-ABC). However, it might still be feasible to proceed to fewer (as compared to ABC) calls to the observation model via its generating process, and acquire a dataset set of recorded observations $\mathcal{D}$.

So in both cases, ABC methods are unfeasible but we can nonetheless dispose of a dataset $\mathcal{D}$. In these situations, statistical learning of the unknown posterior can indeed be an alternative to unfeasible ABC methods. It consists in obtaining an approximation of the unavailable posterior PDF using parametric learning:

$$p_{X|Y=y}(x) \approx p_\theta(x|y); \tag{1.21}$$

where θ is inferred on the dataset $\mathcal{D}$. Then in turn, such an approximation may be used in place of the unavailable posterior distribution in expectations of the form $\mathbb{E}_{\mathcal{P}_{X|Y=y}}[f(X)] \approx \mathbb{E}_{\mathcal{P}_\theta(X|Y=y)}[f(X)]$, by sampling the corresponding model:

$$\mathbb{E}_{\mathcal{P}_{X|Y=y}}[f(X)] \approx \frac{1}{M}\sum_{i=1}^{M} f(x_i), \text{ where } x_i \sim \mathcal{P}_\theta(X|Y=y). \tag{1.22}$$

The principles of approximate Bayesian inference using posterior learning, as opposed to the classical likelihood-based inference, is summarized in Figure 1.1 below.

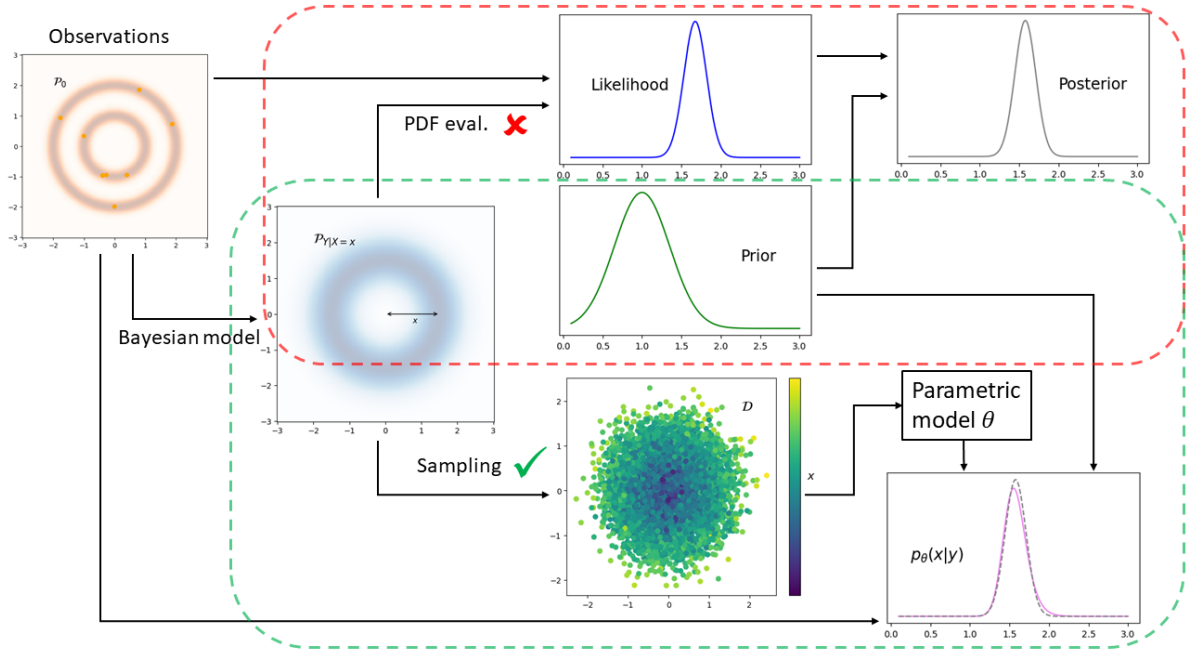


Figure 1.1: Visual summary of posterior learning-based inference versus classical likelihood-based inference

- If the observation model has a tractable PDF, we can indeed compute the likelihood of the observations, which were drawn from $\mathcal{P}_0$. In turn, we can combine the likelihood and the prior to compute the posterior PDF interest, at least up to a constant, as mentioned in previous section 1.1. This corresponds to the usual Bayesian methodology which was recalled in the beginning of this thesis (Northeast part of the picture, red dashed lines).

- However, as we have mentionned, we consider an observation with intractable PDF, making this inference method unfeasible. In that case we can make do of a DGP $\mathcal{P}_{Y|X}$ which is assumed to mimic the unknown process $\mathcal{P}_0$, and from which one can draw $y \sim \mathcal{P}_{Y|X=x}$ for a given input x and thus obtain a dataset comprised of such couples (x, y). This dataset $\mathcal{D}$ enables us to build a parameterized model $\mathcal{P}_\theta(x|y)$ which finally mimics the unknown posterior (Southeast part of the picture, green dashed lines).

1.4 Challenges in Statistical Learning

The previous figure 1.1 presents an illustrative summary of the posterior learning alternative to the classical Bayesian inference approach based on likelihood evaluation. This figure however remains incomplete and lacunar, and raises several questions. There are indeed several aspects of the procedure which cannot be detailed in such an illustration.

A main question that naturally arises is that of the parametric modeling used. In the illustrative figure, we have used the notation $p_\theta(x|y)$, with subscript θ, which implicitly states that the posterior approximation is directly computed via a parametric conditional PDF model. However, this is not necessarily the case and we can indeed approximate the unknown posterior using different modeling approaches.

We first recall that the unavailable posterior PDF is obtained by Bayes formula (3.1) and can be summarized as *prior* times *likelihood* divided by the *evidence*:

$$\underbrace{p_{X|Y=y}(x)}_{posterior} = \frac{\overbrace{p_X(x)}^{prior}\overbrace{p_{Y|X=x}(y)}^{likelihood}}{\underbrace{\int p_X(x)p_{Y|X=x}(y)\mathrm{d}x}_{evidence}}. \tag{1.23}$$

On the one hand, we can suppose that the prior PDF $p_X(x)$ is available for evaluation (this point is discussed in the next 1.4.1). On the other hand, since we have assumed an elaborate observation model (possibly defined as a simulator) with an intractable PDF, the likelihood $p_{Y|X=x}(y)$ cannot be computed. As a consequence, the posterior probability distribution is intractable too, since (i) its numerator involves the unavailable likelihood, and (ii) the evidence at the denominator, which corresponds to the numerator integrated with respect to x, is (doubly) intractable (even if the numerator were indeed tractable, the integral would still not necessarily admit a closed form expression).

So equation (1.23) provides the expression for the target PDF of interest, which is indeed a combination of several factors, with some of them being unavailable for evaluation. We therefore have a choice: we can either approximate the posterior PDF directly as a whole, or we can approximate only its unavailable components. This yields several distinct modeling approaches, which we now describe.

A first possible approach consists in using a parametric model for this posterior PDF. This approach is sometimes referred to as *posterior*, or *discriminative* (D) modeling and is based on the equation:

$$p_{X|Y=y}(x) \overset{D}{\approx} p_\theta(x|y) \tag{1.24}$$

where p_θ is the PDF associated with a probability distribution over X conditioned on the value of RV Y. Examples of this method include the usual methods for classification and regression.

Example: *In classification tasks, we often approximate the probability of classes* $c = 1, ..., C$ *using a model of the form* $\Pr_\theta(X = c|Y = y) = \pi_{c,\theta}(y)$ *where* $[\pi_{1,\theta}(y), ..., \pi_{C,\theta}(y)] = \mathrm{Softmax}(f_\theta(y))$*, and where* $f_\theta(y)$*, which can be a linear function (in which case we talk about logistic classification (33)) or an NN-based function (107), outputs* C *values. The parameters* θ *are often adjusted according to the binary-cross entropy criterion computed using the dataset* $\mathcal{D}$*, which amounts to maximizing the likelihood of* $\mathcal{D}$ *under the model* $\mathcal{P}_\theta(X|Y) = \mathrm{Categorical}(\pi_{1,\theta}(Y), ..., \pi_{C,\theta}(Y))$*. Such a classification model is indeed a discriminative construction since the parametric model directly computes the probabilities (*$\sim$ *PDF) of* X *given observation* Y*.*

Example: *In regression tasks, we often approximate the relationship between* X *and* Y *using a homoskedastic model of the form:* $X = f_\theta(Y) + \sigma\epsilon$*, where* f_θ *is a linear,*

polynomial, or NN-based function, and $\epsilon\sigma \sim \mathcal{N}(0, \sigma^2 I)$ *is a random noise (*σ *may be known or estimated as a parameter). The parameters* θ *(and possibly* σ*) are often adjusted according to the mean squared-error criterion computed using the dataset* $\mathcal{D}$*, which amounts to maximizing the likelihood of* $\mathcal{D}$ *under the model* $\mathcal{P}_\theta(X|Y) = \mathcal{N}(f_\theta(Y), \sigma^2 I)$*. Such a regression model is indeed a discriminative construction since the parametric model directly computes the PDF (via the mean and covariance) of* X *given observation* Y*.*

However, in essence, the posterior PDF has became unavailable mainly because the likelihood function is intractable. So, it is possible to approximate the PDF of the observation model, which yields the following equation:

$$p_{X|Y=y}(x) = \frac{p_X(x)p_{Y|X=x}(y)}{\int p_X(x)p_{Y|X=x}(y)\mathrm{d}x} \overset{G}{\approx} \frac{p_X(x)p_\theta(y|x)}{\int p_X(x)p_\theta(y|x)\mathrm{d}x}, \tag{1.25}$$

where p_θ is the PDF associated with a distribution over Y conditioned on the value of RV X. This approach corresponds to *likelihood* or *generative* (G) modeling.

The discriminative and generative approaches have in common that they both leverage a model of similar structure, which is a conditional probability distribution. However, the former uses this model to compute directly X given Y (see (1.24)) while the latter does the opposite and computes instead Y given X and deduces the corresponding posterior approximation via Bayes formula with the given prior (see (1.25)). In this thesis, more precisely in chapter 3, we propose a comparative study of these two modeling approaches under the scope of epistemic uncertainty quantification.

Finally, in the posterior formula (1.23), as we have mentioned, both the likelihood and the evidence are unavailable, and one can alternatively approximate the Likelihood-To-Evidence Ratio (LTER). This approach is therefore based on the approximation of the unavailable posterior as follows:

$$p_{X|Y=y}(x) = \frac{p_{Y|X=x}(y)}{p_Y(y)} p_X(x) \overset{LTER}{\approx} \rho_\theta(x, y)p_X(x) \tag{1.26}$$

where ρ_θ is a positive function parameterized by θ. Such an approximation can indeed be obtained using a classifier-based PDF ratio approximation (see (60)(36)(95)(96)), which is a popular tool in the statistical and machine learning literature (133)(62)(123)(109)(19). Chapter 2 of this thesis is centered around this approximation method, and it will therefore provide, in its preamble, more details about this learning technique. More precisely, the topic of chapter 2 is related to the task of sampling from such an approximation of the posterior PDF.

In this summarizing figure, we have also eluded the question of the optimization procedure which indeed enables, from (i) a chosen modeling approach (as we have just described) and (ii) a recorded dataset $\mathcal{D}$, to obtain a suitable model.

A pointwise model is often obtained by a learning procedure which minimizes a loss function l:

$$\theta^* = \arg\min_\theta l(\theta, \mathcal{D}). \tag{1.27}$$

In practice, a (possibly approximate) solution to this problem is obtained by reaching a (local) minimum via a gradient-based method. Two notable loss functions are (i) the

negative log-likelihood function $l(\theta, \mathcal{D}) = \log(p(\mathcal{D}|\theta))$, in which case θ^* is the Maximum Likelihood Estimate (MLE) and (ii) the negative log-posterior $l(\theta, \mathcal{D}) = \log(p(\theta|\mathcal{D})) = \log(p(\mathcal{D}|\theta)) + \log(\pi(\theta))$, where the addition of a prior knowledge over θ with $\pi(\theta)$ can be used in practice to induce a regularizing behavior, and yields the maximum a posteriori (MAP) estimate.

On the one hand, for generative and discriminative models, since we are using a conditional probability distribution model (be it over Y given X in the generative case or the opposite in the discriminative case), a tractable PDF of the model is indeed a requirement for computing either the likelihood $p(\mathcal{D}|\theta)$ or the posterior $p(\theta|\mathcal{D})$. It is therefore a relevant problem, in order to use either MLE or MAP training criteria, to consider a parametric model with a tractable PDF. This problem is precisely the topic of chapter 4. In this chapter, we propose a new parametric modeling technique of univariate and conditional distributions such that the corresponding model benefits from straightforward, exact, and fast PDF evaluation.

On the other hand, the LTER approximation of the posterior corresponds to a specific unnormalized (energy-based) model where maximum likelihood is inconvenient since it requires estimating, at each step of a gradient based procedure, the (gradient of) the normalizing constant, which indeed depends on model parameters. In chapter 2, we discuss this precise point, and the learning procedure, i.e. the training of a LTER approximation of the posterior, is presented as an alternative to MLE or MAP which does not require estimating the gradient of the normalizing constant and that instead is defined as a binary classification problem.

In many situations, it is possible that a unique pointwise parameter estimate θ^* does not yield a satisfactory approximating model, even if the loss function is well chosen. This is notably the case in situations where the learning setting can induce an overfitting behavior on the dataset and/or when the considered model is not able to represent the target distribution (this problem also corresponds to a form of misspecification). Bayesian learning proposes an alternative to pointwise estimation methods by instead considering the parameter θ as a RV with prior distribution (as in MAP estimation) and aim marginalizing out this RV in order to compute (or at least to sample from) the posterior predictive distribution (PPD) $p(x|y, \mathcal{D}) = \int p(x, \theta|y, \mathcal{D})\mathrm{d}\theta$. In chapter 3 of this thesis, we explain how to use the PPD in generative and discriminative approaches, and compare the two modeling techniques under the scope of Bayesian learning. In this work, we understand that in order to apply Bayesian learning to either a generative or discriminative model, it is necessary that the corresponding model indeed benefits from a tractable PDF, which further reinforces the importance of chapter 4, where we treat the problem of building a model with a tractable PDF.

At this point, a natural question which arises is that of understanding how different sources of information interact in the inference problem, or, possibly, if they are even taken into account, depending on the modeling choice that we just discussed.

1.4.1 Leveraging prior information in the posterior model

In the previous figure 1.1, the prior distribution is displayed as being related to the posterior approximation. However, in practice, this relationship remains to be precised,

and it is of particular interest to understand how the prior information is involved in the different learning and modeling approaches.

The prior probability distribution describes prior information about the RV of interest X. As we have mentioned before, it is denominated as such since it describes the distribution of that RV *before* having observed that $Y = y$. The prior probability distribution is the result of prior information provided by external sources, transcribed into a prior probability distribution (112) which is sometimes referred to as prior elicitation (14)(94). Obtaining relevant prior information and a relevant prior probability distribution is a scientific or statistical task of its own, which is the topic of several thorough studies (65). We suppose that this prior probability distribution was obtained via an involved procedure of (i) seeking relevant prior information from exogenous sources (for example, from scientific experts), followed by (ii) a transcription procedure to transform this information into a prior distribution. As such, we would ideally not want to discard such information, and we seek a modeling and a learning procedure that indeed accounts for prior information $\mathcal{P}_X$.

It is possible that this prior distribution has a direct impact on the corresponding in the sense that the prior distribution keeps its place in the parameterizing of posterior approximation (see equations (1.25) and (1.26)). Or alternatively, the prior can play an indirect role via the dataset in the case of a discriminative construction, which is one of the points mentioned in the next section and a particular element of interest in chapter 3 of this thesis.

1.4.2 Dataset Acquisition or Augmentation

In figure 1.1, another element which is indeed not detailed, but nonetheless of utmost importance, is that of obtaining the dataset $\mathcal{D}$.

As we have mentioned earlier in section 1.3.3, we suppose that we can dispose of a dataset $\mathcal{D} = \left\{(x_i, y_i)\right\}_{i=1,...,|\mathcal{D}|}$ composed of observed couples $y_i \sim \mathcal{P}_{Y|X=x_i}$ which are each generated by a known input x_i via the observation model of interest. In turn, these couples enable to compensate for the fact that the observation model has an intractable PDF via modeling procedure of the posterior PDF. Indeed, these samples are drawn from (and they provide an empirical approximation of) a joint distribution $\mathcal{P}^{\mathcal{D}}_{X,Y}$ with PDF:

$$p^{\mathcal{D}}_{X,Y}(x, y) = p^{\mathcal{D}}_X(x)p_{Y|X}(y|x); \tag{1.28}$$

where $p^{\mathcal{D}}_X$ is the PDF associated with a given probability distribution which indeed produced the sample values of $x_1, ..., x_{|\mathcal{D}|}$.

Naturally, a main question that arises in statistical modeling is that of constructing the dataset $\mathcal{D}$ appropriately. In most cases, the more recorded couples in the dataset, the better the model represents the true unknown distribution and the more precise the approximate posterior inference. However, in practice, we might have access to a limited budget, and we instead seek to construct the dataset appropriately. So, while in many cases the goal is to best leverage a given dataset at hand, both tasks of (i) constituting a dataset by selecting the distribution $\mathcal{P}^{\mathcal{D}}_X$ and (ii) selecting potentially informative couples to include in the dataset can also be part of the inference and modeling procedure.

Concerning the first point, we stress here that the distribution $\mathcal{P}_X^{\mathcal{D}}$ does not necessarily coincide with the prior $\mathcal{P}_X$ (used to define the posterior distribution of interest in equation (3.1)). This has interesting consequences for the specific discriminative modeling approach (101), which is a point discussed in chapter 3 (see section 3.3.4). In statistical and machine learning, the question of selecting this distribution also has notable connections with topics related to out-of-distribution detection (134), label shift (46), and imbalanced dataset (74).

The second task is mainly covered in the active learning literature (21) and we now distinguish two cases. On the one hand, in some cases, as we mentioned earlier, the observation model is defined via a generative procedure which remains available throughout the learning and inference process, and as such, it is possible to produce observed $y \sim \mathcal{P}_{Y|X=x}$ for a given input x. An operator can therefore acquire new couples (x, y) with which to augment the dataset. Therefore, a natural question that arises is that of determining a relevant (optimal in some sense) value x^* according to an acquisition rule (as in (85)), or designing an appropriate proposal probability distribution from which to draw an x (as in (103)(86)) such that, once associated to a produced y, it constitute a highly informative couple in the learning problem at hand. This problem constitutes a particular case of Bayesian experimental design methods (15)(64)(105) where, again, the PDF of the observation is unavailable to compute an acquisition criteria. Though a consensus has not yet been reached as to which sequential learning scheme seems to be the best, it is clear from empirical studies that sequential methods based on motivated acquisition rules produce efficient estimation of model parameters, leading to improved inference usually while reducing the number of calls to the simulation mechanism of the observation model (37)(36)(119).

On the other hand, however, in many learning settings, the dataset is built in reverse via a *labeling* procedure, which we describe. The y-values are samples drawn from $\mathcal{P}_Y^{\mathcal{D}}(\mathrm{d}y) = \int \mathcal{P}_{Y|X}(\mathrm{d}y|x)\mathcal{P}_X^{\mathcal{D}}(\mathrm{d}x)$ but without having recorded the value of an associated x_i. Then, an oracle (in many cases a human operator, a scientific expert, or an otherwise computationally intensive operation) is called upon to associate to y its corresponding value of x. In this case, the question of suitably creating the dataset $\mathcal{D}$ usually involves selecting a value y among a set of unlabeled observations, usually referred to as a pool, for which to call upon the oracle (45) to obtain the corresponding x, such that once this couple is used to augment the dataset, the resulting modeling becomes increasingly accurate.

1.4.3 Inference from multiple observations acting as unlabeled dataset

Thus far, we have motivated the interest of modeling the unknown posterior for predicting the value of a RV X given that we observe the value of $Y = y$ from a random process of interest $\mathcal{P}_0$. However, this situation corresponds to a specific formulation of the inference problem, and, in practice we might instead dispose of multiple observations, and this multiplicity can have different meanings which we now explain.

Firstly, it is possible that we wish to make predictions for different couples of RVs (X_i, Y_i). More precisely, we might observe the values of a set of RVs $\{Y_i \sim \mathcal{P}_i\}$ where

all $\mathcal{P}_i$ are random processes of interest, which we believe can efficiently be described via the same observation model $\mathcal{P}_{Y|X}$ but for different underlying RVs X_i. In the context of an implicit observation model, we therefore might leverage a common model to predict all the RVs X_i given all the observations $Y_i = y_i$.

Secondly, it is possible that we dispose of similar observations $Y_i = y_i$, each from a random process $\mathcal{P}_i$ which, we indeed believe, as before, can be efficiently modeled using a same observation model $\mathcal{P}_{Y|X}$, but that the random process is not of particular interest in the scientific problem, such that we do not consider the goal of inferring an associated underlying RV X_i.

Thirdly, it is possible that we dispose of recorded values $Y_i = y_i$ from the observation model without having explicitly recorded its corresponding input value x_i.

In all these three settings, even though the values x_i associated with an observed $Y_i = y_i$ are not necessarily of interest, it turns out that all these observations can bring information in the modeling problem. Indeed, the observations are either produced by random phenomena, which we assumed could be accurately represented by the observation model, or by the observation model itself. So, in addition to the dataset $\mathcal{D}$, the corresponding parametric model, which accounts for the intractable PDF of the observation model, should indeed be in coherence with these observations, which can indeed be understood as unlabeled observations. In these settings, we thus obtain a semi-supervised learning problem.

Lastly, it is possible that, from a given random process of interest, say $\mathcal{P}_0$, we observe the value of several iid RVs $Y_{0,1} = y_{0,1}, ..., Y_{0,N_0} = y_{0,N_0}$; and studying the random process of interest amounts to inferring the value of a RV of interest X_0 which can be related to the iid RVs $Y_{0,1}, ..., Y_{0,N_0}$ via the observation model.

These questions are topics of interest covered in the chapter 3 of this thesis where we notably compare the ability of generative and discriminative modeling approaches to infer from multiple observations in these different cases.

1.5 Conclusion

We now summarize the discussion of this chapter, which explains the different reasons and situations which lead practitioners to progressively turn a Bayesian inference problem into a problem of statistical learning nature.

We first recalled the principles of Bayesian posterior inference. This methodology enables to interpret and understand phenomena of nature by studying the underlying probable causes. By selecting a variable of interest X which can be related to an observation Y, and by specifying a joint distribution over this couple of RVs, we can retrieve the probable causes with the posterior distribution.

We explained how to proceed with inference in practice, by covering the usual Monte-Carlo estimation methods. By reviewing the most common algorithms for sampling from a distribution, we emphasized the central role played by the (posterior) PDF.

We also interpreted Bayesian inference as a way to obtain a probabilistic model of a phenomenon of interest using an observation model. We stressed the importance of an accurate observation model and discussed its possible misspecification. In many

scientific fields, advances through research, studies, and experiments are leading practitioners to consider ever more accurate and intricate representations of nature, and pave the way to increasingly complex models. However, when using such models for Bayesian inference, we can face the major drawback that the corresponding observation model no longer benefits from a tractable PDF, making usual posterior inference unfeasible since the posterior PDF then suffers from the same intractability.

This finally led us to statistical learning, which is a way to cope with this shortcoming as it indeed enables performing approximate posterior inference in the case where the observation model has an intractable PDF (and thus the posterior too). Since the classical methods for statistical learning usually require a dataset to obtain such an approximation, learning-based approximate Bayesian inference is particularly well-suited (and has already been successfully applied) in the case where the observation model is defined as a simulator. Finally, using a generic figure which summarizes the statistical learning methodology as an alternative to usual Bayesian inference, we identified several challenges, some of which are going to be discussed in detail in the following chapters of this thesis.

Bibliography

[1] Sergios Agapiou, Omiros Papaspiliopoulos, Daniel Sanz-Alonso, and Andrew M Stuart. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, pages 405–431, 2017.

[2] Anthony Alfred Barker. Monte carlo calculations of the radial distribution functions for a proton electron plasma. *Australian Journal of Physics*, 18(2):119–134, 1965.

[3] Thomas Bayes. LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philosophical transactions of the Royal Society of London*, (53):370–418, 1763.

[4] José M Bernardo and Adrian FM Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

[5] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[6] George EP Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

[7] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov chain Monte Carlo*. CRC press, 2011.

[8] O. Cappé, É. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer-Verlag, 2005.

[9] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer Series in Statistics. Springer, 2005.

[10] Olivier Cappé, Randal Douc, Arnaud Guillin, Jean-Michel Marin, and Christian P Robert. Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18:447–459, 2008.

[11] Olivier Cappé, Arnaud Guillin, Jean-Michel Marin, and Christian P Robert. Population monte carlo. *Journal of Computational and Graphical Statistics*, 13(4):907–929, 2004.

[12] George Casella, Christian P Robert, and Martin T Wells. Generalized accept-reject sampling schemes. *Lecture Notes-Monograph Series*, pages 342–347, 2004.

[13] Anthony Caterini, Rob Cornish, Dino Sejdinovic, and Arnaud Doucet. Variational inference with continuously-indexed normalizing flows. In *Uncertainty in Artificial Intelligence*, pages 44–53. PMLR, 2021.

[14] Kathryn Chaloner. Elicitation of prior distributions. *Bayesian biostatistics*, 141:156, 1996.

[15] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical science*, pages 273–304, 1995.

[16] Sourav Chatterjee and Persi Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135, 2018.

[17] Ming-Hui Chen and Joseph G Ibrahim. Conjugate priors for generalized linear models. *Statistica Sinica*, pages 461–476, 2003.

[18] Siddhartha Chib and Edward Greenberg. Understanding the Metropolis-Hastings algorithm. *The american statistician*, 49(4):327–335, 1995.

[19] Kristy Choi, Madeline Liao, and Stefano Ermon. Featurized density ratio estimation. In *Uncertainty in Artificial Intelligence*, pages 172–182. PMLR, 2021.

[20] Grégoire Clarté, Christian P Robert, Robin J Ryder, and Julien Stoehr. Componentwise approximate Bayesian computation via Gibbs-like steps. *Biometrika*, 108(3):591–607, 2021.

[21] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996.

[22] Adrien Corenflos, James Thornton, George Deligiannidis, and Arnaud Doucet. Differentiable particle filtering via entropy-regularized optimal transport. In *International Conference on Machine Learning*, pages 2100–2111. PMLR, 2021.

[23] SR Dalal and WJ Hall. Approximating priors by mixtures of natural conjugate priors. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):278–286, 1983.

[24] Paul Damlen, John Wakefield, and Stephen Walker. Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):331–344, 1999.

[25] Philip J Davis and Philip Rabinowitz. *Methods of numerical integration*. Courier Corporation, 2007.

[26] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

[27] Luc Devroye. Sample-based non-uniform random variate generation. In *Proceedings of the 18th conference on Winter simulation*, pages 260–265, 1986.

[28] Luc Devroye. Nonuniform random variate generation. *Handbooks in operations research and management science*, 13:83–121, 2006.

[29] Akash Kumar Dhaka, Alejandro Catalina, Manushi Welandawe, Michael R Andersen, Jonathan Huggins, and Aki Vehtari. Challenges and opportunities in high dimensional variational inference. *Advances in Neural Information Processing Systems*, 34:7787–7798, 2021.

[30] Persi Diaconis and Donald Ylvisaker. Conjugate priors for exponential families. *The Annals of statistics*, pages 269–281, 1979.

[31] Peter J Diggle and Richard J Gratton. Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 46(2):193–212, 1984.

[32] Randal Douc, Arnaud Guillin, J-M Marin, and Christian P Robert. Convergence of adaptive mixtures of importance sampling schemes. 2007.

[33] Stephan Dreiseitl and Lucila Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 35(5-6):352–359, 2002.

[34] Christopher C Drovandi and Anthony N Pettitt. Bayesian experimental design for models with intractable likelihoods. *Biometrics*, 69(4):937–948, 2013.

[35] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

[36] Conor Durkan, Iain Murray, and George Papamakarios. On contrastive learning for likelihood-free inference. In *International conference on machine learning*, pages 2771–2781. PMLR, 2020.

[37] Conor Durkan, George Papamakarios, and Iain Murray. Sequential neural methods for likelihood-free inference. *arXiv preprint arXiv:1811.08723*, 2018.

[38] Richard G Everitt, Adam M Johansen, Ellen Rowing, and Melina Evdemon-Hogan. Bayesian model comparison with intractable likelihoods. *arXiv preprint arXiv*, 1504(06697664):10–1007, 2015.

[39] Paul Fearnhead and Dennis Prangle. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 74(3):419–474, 2012.

[40] Mikhail Figurnov, Shakir Mohamed, and Andriy Mnih. Implicit reparameterization gradients. *Advances in neural information processing systems*, 31, 2018.

[41] Daniel Fink. A compendium of conjugate priors. *See http://www. people. cornell. edu/pages/df36/CONJINTRnew% 20TEX. pdf*, 46, 1997.

[42] Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368, 1922.

[43] Bernard D Flury. Acceptance–rejection sampling made easy. *Siam Review*, 32(3):474–476, 1990.

[44] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[45] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR, 2017.

[46] Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary Lipton. A unified view of label shift estimation. *Advances in Neural Information Processing Systems*, 33:3290–3300, 2020.

[47] Tomas Geffner and Justin Domke. On the difficulty of unbiased alpha divergence minimization. *arXiv preprint arXiv:2010.09541*, 2020.

[48] Alan E Gelfand. Gibbs sampling. *Journal of the American statistical Association*, 95(452):1300–1304, 2000.

[49] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[50] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

[51] James Gentle, Wolfgang Karl Härdle, and Yuichi Mori. *Handbook of Computational Statistics: Concepts and Methods*. Springer Berlin Heidelberg, 2012.

[52] James E Gentle. *Random number generation and Monte Carlo methods*, volume 381. Springer, 2003.

[53] John Geweke. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica: Journal of the Econometric Society*, pages 1317–1339, 1989.

[54] Walter R Gilks and Pascal Wild. Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(2):337–348, 1992.

[55] Gerald Goertzel. Quota sampling and importance functions in stochastic solution of particle problems. Technical report, 1949.

[56] Christian Gourieroux and Alain Monfort. *Simulation-based econometric methods*. Oxford university press, 1996.

[57] Ulf Grenander and Michael I Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(4):549–581, 1994.

[58] John Michael Hammersley and David Christopher Handscomb. General principles of the Monte Carlo method. In *Monte Carlo Methods*, pages 50–75. Springer, 1964.

[59] W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. 1970.

[60] Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free mcmc with amortized approximate ratio estimators. In *International conference on machine learning*, pages 4239–4248. PMLR, 2020.

[61] Timothy Classen Hesterberg. *Advances in importance sampling*. Stanford University, 1988.

[62] Shohei Hido, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, and Takafumi Kanamori. Statistical outlier detection using direct density ratio estimation. *Knowledge and information systems*, 26:309–336, 2011.

[63] Matthew D Hoffman, Andrew Gelman, et al. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[64] Xun Huan and Youssef M Marzouk. Simulation-based optimal Bayesian experimental design for nonlinear systems. *Journal of Computational Physics*, 232(1):288–317, 2013.

[65] Edwin T Jaynes. Prior probabilities. *IEEE Transactions on systems science and cybernetics*, 4(3):227–241, 1968.

[66] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Learning in graphical models*, pages 105–161, 1998.

[67] H. Kahn and A. W. Marshall. Methods of Reducing Sample Size in Monte Carlo Computations. *J. of the Op. Res. Soc. of Amer.*, 1(5):263–278, 1953.

[68] Herman Kahn. Stochastic (monte carlo) attenuation analysis. *RAND Corporation Report R-163, The RAND Corporation, Santa Monica, Calif*, 1949.

[69] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[70] Albert J Kinderman and John F Monahan. Computer generation of random variables using the ratio of uniform deviates. *ACM Transactions on Mathematical Software (TOMS)*, 3(3):257–260, 1977.

[71] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[72] Diederik P Kingma and Max Welling. Auto-Encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[73] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29, 2016.

[74] Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, et al. Handling imbalanced datasets: A review. *GESTS international transactions on computer science and engineering*, 30(1):25–36, 2006.

[75] Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

[76] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[77] Roland Lamberti, Yohan Petetin, François Septier, and François Desbouvries. A double proposal normalized importance sampling estimator. In *2018 IEEE Statistical Signal Processing Workshop (SSP)*, pages 238–242. IEEE, 2018.

[78] Kenneth Lange, J Chambers, and W Eddy. *Numerical analysis for statisticians*, volume 1. Springer, 2010.

[79] Krzysztof Łatuszyński and Gareth O Roberts. CLTs and asymptotic variance of time-sampled Markov chains. *Methodology and Computing in Applied Probability*, 15:237–247, 2013.

[80] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[81] Yingzhen Li and Richard E Turner. Rényi divergence variational inference. *Advances in neural information processing systems*, 29, 2016.

[82] Jun S Liu and Jun S Liu. *Monte Carlo strategies in scientific computing*, volume 75. Springer, 2001.

[83] Samuel Livingstone and Giacomo Zanella. The Barker proposal: combining robustness and efficiency in gradient-based MCMC. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):496–523, 2022.

[84] Alfred James Lotka. *Elements of physical biology*. Williams & Wilkins, 1925.

[85] Jan-Matthis Lueckmann, Giacomo Bassetto, Theofanis Karaletsos, and Jakob H Macke. Likelihood-free inference with emulator networks. In *Symposium on Advances in Approximate Bayesian Inference*, pages 32–53. PMLR, 2019.

[86] Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnenmacher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural dynamics. *Advances in neural information processing systems*, 30, 2017.

[87] Anne-Marie Lyne, Mark Girolami, Yves Atchadé, Heiko Strathmann, and Daniel Simpson. On Russian roulette estimates for Bayesian inference with doubly-intractable likelihoods. 2015.

[88] Jean-Michel Marin, Pierre Pudlo, Christian P Robert, and Robin J Ryder. Approximate Bayesian computational methods. *Statistics and computing*, 22(6):1167–1180, 2012.

[89] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.

[90] Paul Marjoram and Simon Tavaré. Modern computational approaches for analysing molecular genetic variation data. *Nature Reviews Genetics*, 7(10):759–770, 2006.

[91] A. W. Marshall. The use of multi-stage sampling schemes in Monte Carlo computations. In M. Meyer, editor, *Symposium on Monte Carlo Methods*, pages 123–140, New York, 1956.

[92] Luca Martino, David Luengo, and Joaquín Míguez. *Independent random sampling methods*. Springer, 2018.

[93] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[94] Petrus Mikkola, Osvaldo A Martin, Suyog Chandramouli, Marcelo Hartmann, Oriol Abril Pla, Owen Thomas, Henri Pesonen, Jukka Corander, Aki Vehtari, Samuel Kaski, et al. Prior knowledge elicitation: The past, present, and future. *arXiv preprint arXiv:2112.01380*, 2021.

[95] Benjamin K Miller, Alex Cole, Patrick Forré, Gilles Louppe, and Christoph Weniger. Truncated marginal neural ratio estimation. *Advances in Neural Information Processing Systems*, 34:129–143, 2021.

[96] Benjamin K Miller, Christoph Weniger, and Patrick Forré. Contrastive neural ratio estimation. *Advances in Neural Information Processing Systems*, 35:3262–3278, 2022.

[97] Christian Naesseth, Francisco Ruiz, Scott Linderman, and David Blei. Reparameterization gradients through acceptance-rejection sampling algorithms. In *Artificial Intelligence and Statistics*, pages 489–498. PMLR, 2017.

[98] Radford Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 06 2012.

[99] Radford M Neal. Slice sampling. *The annals of statistics*, 31(3):705–767, 2003.

[100] Art B Owen. Monte Carlo theory, methods and examples, 2013.

[101] George Papamakarios and Iain Murray. Fast ε-free inference of simulation models with bayesian conditional density estimation. *Advances in neural information processing systems*, 29, 2016.

[102] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):2617–2680, 2021.

[103] George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 837–848. PMLR, 2019.

[104] Jaewoo Park and Murali Haran. Bayesian inference in the presence of intractable normalizing functions. *Journal of the American Statistical Association*, 113(523):1372–1390, 2018.

[105] Tom Rainforth, Adam Foster, Desi R Ivanova, and Freddie Bickford Smith. Modern bayesian experimental design. *arXiv preprint arXiv:2302.14545*, 2023.

[106] Oliver Ratmann, Christophe Andrieu, Carsten Wiuf, and Sylvia Richardson. Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proceedings of the National Academy of Sciences*, 106(26):10576–10581, 2009.

[107] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017.

[108] Sebastian Reich. A nonparametric ensemble transform method for Bayesian inference. *SIAM Journal on Scientific Computing*, 35(4):A2013–A2024, 2013.

[109] Benjamin Rhodes, Kai Xu, and Michael U Gutmann. Telescoping density-ratio estimation. *Advances in neural information processing systems*, 33:4905–4916, 2020.

[110] Steven Riley, Christophe Fraser, Christl A Donnelly, Azra C Ghani, Laith J Abu-Raddad, Anthony J Hedley, Gabriel M Leung, Lai-Ming Ho, Tai-Hing Lam, Thuan Q Thach, et al. Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. *Science*, 300(5627):1961–1966, 2003.

[111] Oren Rippel and Ryan Prescott Adams. High-dimensional probability estimation with deep density models. *arXiv preprint arXiv:1302.5125*, 2013.

[112] Christian P. Robert. *From Prior Information to Prior Distributions*, chapter Chapter Number, pages 89–135. Springer New York, New York, NY, 1994.

[113] Christian P Robert and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

[114] Christian P Robert et al. *The Bayesian choice: from decision-theoretic foundations to computational implementation*, volume 2. Springer, 2007.

[115] Gareth O Roberts and Osnat Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and computing in applied probability*, 4:337–357, 2002.

[116] Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

[117] D. B. Rubin. Using the SIR algorithm to simulate posterior distributions. In *Bayesian Statistics III*. Oxford University Press, Oxford, 1988.

[118] Donald B Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, pages 1151–1172, 1984.

[119] Louis Sharrock, Jack Simons, Song Liu, and Mark Beaumont. Sequential neural score estimation: Likelihood-free inference with conditional score based diffusion models. *arXiv preprint arXiv:2210.04872*, 2022.

[120] Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of approximate Bayesian computation*. CRC Press, 2018.

[121] Scott A Sisson, Yanan Fan, and Mark M Tanaka. Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.

[122] Adrian FM Smith and Alan E Gelfand. Bayesian statistics without tears: a sampling–resampling perspective. *The American Statistician*, 46(2):84–88, 1992.

[123] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64:1009–1044, 2012.

[124] Simon Tavaré, David J Balding, Robert C Griffiths, and Peter Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518, 1997.

[125] Achille Thin, Nikita Kotelevskii, Jean-Stanislas Denain, Leo Grinsztajn, Alain Durmus, Maxim Panov, and Eric Moulines. MetFlow: a new efficient method for bridging the gap between Markov chain Monte Carlo and variational inference. *arXiv preprint arXiv:2002.12253*, 2020.

[126] Luke Tierney. Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728, 1994.

[127] Vito Volterra. Fluctuations in the abundance of a species considered mathematically. *Nature*, 119(2983):12–13, 1927.

[128] John Von Neumann. Various techniques used in connection with random digits. *John von Neumann, Collected Works*, 5:768–770, 1963.

[129] Dilin Wang, Hao Liu, and Qiang Liu. Variational inference with tail-adaptive f-divergence. *Advances in Neural Information Processing Systems*, 31, 2018.

[130] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

[131] Simon N Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104, 2010.

[132] Ming Xu, Matias Quiroz, Robert Kohn, and Scott A Sisson. Variance reduction properties of the reparameterization trick. In *The 22nd international conference on artificial intelligence and statistics*, pages 2711–2720. PMLR, 2019.

[133] Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Advances in neural information processing systems*, 24, 2011.

[134] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, pages 1–28, 2024.

[135] David Zoltowski, Diana Cai, and Ryan P Adams. Slice Sampling Reparameterization Gradients. *Advances in Neural Information Processing Systems*, 34:23532–23544, 2021.

Chapter 2

Likelihood-to-evidence ratio posterior sampling

As we have mentioned in the introduction of this thesis, a possible approach to obtain a parametric approximation of the unknown posterior of interest is the method of approximating the likelihood-to-evidence ratio (LTER). This method was proposed in (55) and it has become a prevalent method for likelihood-free inference (18)(28)(56). The starting point of this principle is that the unknown posterior PDF can be rewritten as:

$$p_{X|Y=y}(x) = \frac{p_{Y|X=x}(y)p_X(x)}{p_Y(y)p_X(x)}p_X(x); \tag{2.1}$$

where, we recall, $p_Y(y) = \int p_{Y|X=x}(y)p_X(x)\mathrm{d}x$ is the evidence. So, we see that the first factor, which after simplification of the prior PDF is nothing but the LTER, can also be seen as a ratio between two joint PDFs: the numerator $p_{Y|X=x}(y)p_X(x)$ is the PDF of the joint distribution $\mathcal{P}_{X,Y}$ while in the denominator $p_Y(y)p_X(x)$ the product of marginal and can be seen as the PDF of $\mathcal{P}_X \otimes \mathcal{P}_Y$. As is well known and well established in the literature (6)(52), one can leverage a classifier which is designed to distinguish samples from the two probability distributions by approximating an underlying class posterior in an implicit binary mixture context to obtain an approximation of the probability density function (PDF) ratio. Let us denote $r_\theta(x, y)$ such a classifier; the corresponding approximation of the unknown posterior reads:

$$p_{X|Y=y}(x) \approx \frac{r_\theta(x, y)}{1 - r_\theta(x, y)}p_X(x). \tag{2.2}$$

In this approximation, $r_\theta \in [0, 1]$ is a probability, which is the output of a binary classification function parameterized by θ. This function is obtained by adjusting the parameters according to the Binary Cross-Entropy criterion:

$$\mathcal{L}_{\mathrm{BCE}}(\theta) = -\mathbb{E}_{\mathcal{P}_{X,Y}}\big[\log(r_\theta(X, Y))\big] - \mathbb{E}_{\mathcal{P}_Y \otimes \mathcal{P}_X}\big[\log(1 - r_\theta(X, Y))\big]; \tag{2.3}$$

where both expectations can be estimated using, respectively, the dataset $\mathcal{D}$ and a shuffled version of $\mathcal{D}$.

In this context, the scope of this contribution is related to the question of sampling from the distribution associated with this approximation. Indeed, equation (2.2)

provides an (approximately normalized) model for the posterior PDF, which can be computed up to a normalizing constant. Therefore, any MCMC scheme can be applied to sample from the underlying probability distribution as proposed in (28). However, designing an efficient MCMC scheme requires fine-tuning the transition kernel in order to avoid slow mixing of the Markov chain. With that regard, we instead propose to leverage the structure of the LTER approximation of the posterior PDF (2.2) which we explain is compatible with the three specific MC sampling algorithms of Accept-Reject (AR), Independent-Metropolis-Hastings (IMH) and Importance Sampling (IS) with Sampling Importance Resampling (SIR).

The scope of our work is therefore centered around the problem of sampling from a distribution $\mathcal{P}_\theta$ which is a ratio-based approximation of a distribution of interest $\mathcal{P}$ (in this context, the posterior distribution $\mathcal{P}_{X|Y=y}$). Our approach turns this problem around: we propose to perform approximate sampling from $\mathcal{P}$ using classical ratio-involved Monte Carlo (MC) sampling techniques of AR, IMH and IS where the unknown PDF ratio is instead replaced by an approximation based on a classifier trained to distinguish from a given instrumental distribution (in this context, the prior distribution $\mathcal{P}_X$).

2.1 Binary Classification based Monte Carlo sampling

AR, IMH or IS MC algorithms all involve computing *ratios* of two PDFs p_1 and p_0. On the other hand, classifiers discriminate samples produced by a binary mixture and can be used to approximate the *ratio* of corresponding PDFs. We therefore establish a bridge between simulation and classification, which enables us to propose PDF-free versions of ratio-based simulation algorithms, where the ratio is replaced by a surrogate function computed via a classifier. Our modified samplers are based on very different hypotheses: the knowledge of functions p_1 and p_0 is relaxed (- they may be totally unknown), and is counterbalanced by the availability of a classification function, which can be obtained from a labeled dataset. From a probabilistic modeling perspective, our procedure involves a structured energy based model (EBM) which can easily be trained and is structurally compatible with the classical samplers.

2.1.1 Introduction

If a and b are two positive numbers,

$$r = \frac{a}{a+b} \in (0,1) \Leftrightarrow \frac{r}{1-r} = \frac{a}{b} > 0. \tag{2.4}$$

This equivalence has interesting consequences in Bayesian classification, machine learning and stochastic simulation. Indeed, if a and b are probabilities of two classes in a binary mixture context for a given sample, then ratio $\frac{a}{a+b}$ is the posterior probability which provides with the class probabilities for a given sample, and can be approximated

by a parametric classifier r_θ trained to distinguish between the two probability distributions. On the other hand, positive ratios $\frac{a}{b}$ play a key role in AR, IMH or IS techniques. Equation (2.4) relates r to such positive ratios, and tells us that ratio $\frac{a}{b}$ can be computed exactly from r, or, in practice, approximately from r_θ, without necessarily knowing a nor b. This observation enables us to propose approximate versions of these algorithms which rely on weaker hypotheses.

Let $\lambda, 1-\lambda \in (0,1)$ be the prior probabilities of two categories $k = 1, 0$, distributed resp. $\sim p_1$ and p_0. Binary classification distinguishes samples from mixture $\lambda p_1 + (1-\lambda)p_0$ by identifying the PDF which generated them. The appropriate way to classify relies on the posterior probability: x is a sample $\sim p_1$ rather than $\sim p_0$ with probability

$$\Pr(k=1|x,\lambda,p_0,p_1) = \frac{\lambda p_1(x)}{\lambda p_1(x) + (1-\lambda)p_0(x)}. \tag{2.5}$$

Indeed, as is well known (see e.g. (22, Chap. 11)), assigning a sample to the label with highest posterior probability is the optimal decision rule in the sense that it minimizes the probability of misclassification.

To compute this posterior probability, one needs to evaluate the PDFs p_1, p_0 and know the prior probability λ but they are often unknown, leaving (2.5) intractable. If however we dispose of a set $\mathcal{D} = \{(x_i^{(k_i)}, k_i)\}_{i=1}^{N_0+N_1}$ of labelled observations, we can make use of a parametric classifier. So let us assume that we have at our disposal a classifier function r_θ, parameterized by θ which mimics the unknown posterior PDF:

$$r_\theta(x) \approx \frac{N_1 p_1(x)}{N_1 p_1(x) + N_0 p_0(x)}. \tag{2.6}$$

Our approach is based on the observation that (2.6) is equivalent to

$$\frac{N_0}{N_1}\frac{r_\theta(x)}{1-r_\theta(x)} \approx \frac{p_1(x)}{p_0(x)}, \tag{2.7}$$

which implies that (typically neural network (NN)-based) classifiers can also be used for approximating *PDF ratios*.

Equation (2.7) has already been observed, and exploited in contexts where estimating a ratio of PDFs is relevant. First, classifiers are at the core of adversarial training techniques in which divergence measures involving a ratio are replaced by an approximation based on a classifier (42). This enables learning implicit generative models (i.e., with intractable PDFs) (23) (14). Moreover, classifier-based PDF ratio approximation has been applied to estimation of such metrics as Mutual Information (5). Finally, classifiers based ratios have been applied successfully in statistical hypothesis testing procedures (25), which heavily rely on likelihood-ratio tests.

If p_0 is an instrumental distribution with a tractable PDF, then (2.7) can be easily be turned into an approximation of target PDF p_1. So classifiers can be used for density estimation, conditional density estimation, or LTER estimation, making them especially relevant in a likelihood-free inference setting (18)(28)(56).

However, the question of *sampling* from the corresponding model remains open, and this is precisely the point we discuss in this chapter. We realize that PDF ratios

also play a key role in such simulation techniques as the AR or Markov Chain Monte Carlo (MCMC) methods, in which samples from instrumental p_0 are transformed into samples from the target p_1 via a sampling mechanism which involves the ratio of the two densities. This establishes a connection between classification and MC sampling, and will enable us to relax the assumption of tractable PDF p_0, p_1 of these sampling algorithms, at the price of approximate sampling. Our approach is therefore completely PDF-free, and as such is especially relevant when the target distribution is unknown or with intractable, noisy, or costly to evaluate PDF (see (40) for a review of MC techniques in this setting, and (48) for a review of likelihood-free Approximate Bayesian Computation techniques); and/or when the instrumental p_0 is defined by a generative model with implicit PDF (36)(23)(50). The rest of this section is organized as follows. In §2.1.2 we recall classical ratio-based stochastic simulation algorithms, i.e. the AR, IMH and IS techniques. In §2.1.3 we show that classifiers computed via the Binary Cross Entropy (BCE) criterion indeed provide with an approximation of the posterior (2.5). Finally in §2.1.4 we propose classification based sampling methods, illustrate our method via simulations[1], and revisit it under the perspective of probabilistic modeling.

2.1.2 Classical ratio-based sampling algorithms

Stochastic simulation includes a variety of techniques, see e.g. (2)(20). In this section we focus on AR, IMH and IS which share in common that they all compute a ratio of PDFs.

AR

AR Sampling (46, chap. 2) (40, chap. 3) is a simulation algorithm that yields samples distributed according to a target distribution p via samples from a proposal distribution q, which are accepted or rejected as valid samples from p via some acceptance probability. More precisely, let the support of p be inside that of q. This means that there exists a constant $C \geq 1$ such that for all $x \in \mathbb{R}^d, p(x) \leq Cq(x)$. Let $X \sim q$, and let k a Bernoulli random variable with parameter $\alpha_{AR}(X) = \frac{p(X)}{Cq(X)}$. AR sampling is based on the fact that $X|k = 1$ is distributed according to p. Note that $\Pr(k = 1) = \frac{1}{C}$, so the lower the value of C, the higher the acceptance rate.

In order to use the algorithm in practice, we thus need to be able to evaluate PDF p, and build q such that one can sample easily from q and there exists C such that $p(x) \leq Cq(x)$ for all x, we can compute one such value of C, and C is as small as possible. Note finally that the algorithm can easily be adapted to the cases where p and/or q are known up to a (non necessarily common) constant, see e.g. (43, Th. 4.5).

As we shall now see, AR sampling is indeed nothing but a binary classification procedure (see also (10, §6) for an application of this principle).

Starting from the target PDF $p(x)$, we find an easy-to-sample distribution Q and constant $C > 1$ such that $Cq(x)$ envelopes $p(x)$. Since $Cq(x) - p(x)$ is non negative, we write $Cq(x)$ as $p(x)$ plus a positive remainder which, up to a constant, is also a PDF;

[1] Code available at github.com/ElouanARGOUARCH/Binary-Classification-Based-Monte-Carlo-Simulation

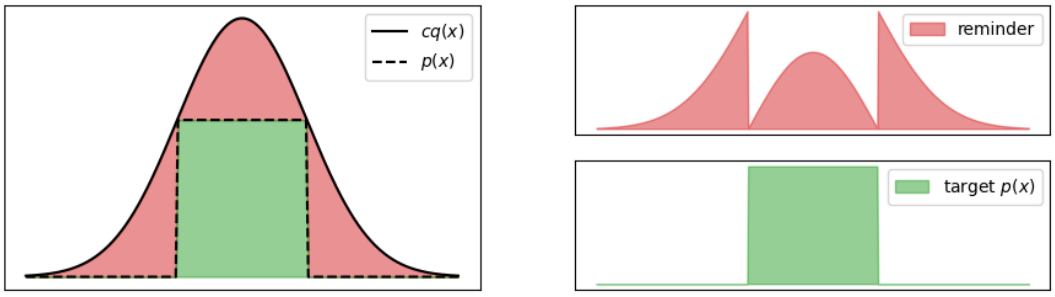


Figure 2.1: Enveloping target PDF builds an implicit mixture.

so enveloping $p(x)$ with $Cq(x)$ is nothing but building the implicit binary mixture PDF (see also fig. 2.1 below)

$$\underbrace{q(x)}_{\text{proposal}} = \frac{1}{C}\underbrace{p(x)}_{\text{target}} + (1-\frac{1}{C})\underbrace{\frac{q(x)-\frac{1}{C}p(x)}{1-\frac{1}{C}}}_{\text{remainder}} \tag{2.8}$$

with a priori probabilities $\frac{1}{C}$ and $1-\frac{1}{C}$. The first component of the mixture is the target PDF p, and the second one is the PDF of the rejected samples. The score point of AR consists in drawing samples from q *without* needing to sample from its two implicit mixture components (see the r.h.s. of (2.8)). Accepting (or rejecting) a sample depending on the ratio probability

$$\alpha_{AR}(x) = \frac{p(x)}{Cq(x)} = \frac{\frac{1}{C}p(x)}{\frac{1}{C}p(x) + (1-\frac{1}{C})\frac{q(x)-\frac{1}{C}p(x)}{1-\frac{1}{C}}} \tag{2.9}$$

then amounts to classifying the samples with the posterior PDF (compare (2.9) to (2.5)).

IMH

MCMC algorithms build a Markov chain whose invariant distribution is the target distribution p; so simulating the chain yields samples asymptotically distributed $\sim p$. The Metropolis-Hastings (MH) algorithm (46) (12) is a particular MCMC method in which the transition is a two-step procedure: given a current state x_t, we propose x^* from $q(.|x_t)$, and then we compute the acceptance probability $\alpha_{MH}(x^*, x_t) = \min(1, \frac{p(x^*)q(x_t|x^*)}{p(x_t)q(x^*|x_t)})$. x^* is accepted as the new state x_{t+1} with probability $\alpha_{MH}(x^*, x_t)$; if x^* is rejected then the chain remains in the current state x_t. In practice, $q(.|.)$ plays a crucial role in the performance of the MH algorithm: if not well-tuned, it can lead to a poor exploration of the target distribution.

The IMH algorithm is a simplified version of MH which considers an independent transition. The new point x^* is hence proposed independently of the current state x_t, according to an independent proposal $q(.)$. In this case, the acceptance probability reduces to $\alpha_{IMH}(x^*, x_t) = \min(1, \frac{p(x^*)q(x_t)}{p(x_t)q(x^*)})$.

IS

Many problems involve computing the expectation of some function f with respect to PDF p: $\mathbb{E}_p\big[f(x)\big] = \int f(x)p(x)\mathrm{d}x$. In practice the integral can be intractable, so we may need to resort to MC approximations. IS is a technique for reducing the variance of such MC estimates which can be traced back to the 1950's (35) (39) (27, §5.4).

The crude MC estimate reads $\frac{1}{N}\sum_{i=1}^N f(x_i), x_i \overset{\text{iid}}{\sim} p$. However, on the one hand it is generally difficult to sample directly from p, and on the other hand it can yield a poor estimate when the regions where p is large do not coincide with those where f is large. Rewriting $\mathbb{E}_p\big[f(x)\big] = \mathbb{E}_q\left[\frac{p(x)f(x)}{q(x)}\right]$, for some importance distribution q, leads to the IS estimate $\frac{1}{N}\sum_{i=1}^N \frac{p(x_i)}{q(x_i)}f(x_i), x_i \overset{\text{iid.}}{\sim} q$. One can easily show that the PDF which minimizes the variance is, up to a constant, $|f(x)|p(x)$. Even if this optimal importance distribution cannot be used in practice, this tells us that the regions of *importance* are not those where p is large, but rather those where $|f|p$ is large. Note that the IS estimate can be computed only if p and q are known exactly, or known up to a common constant; if this is not the case one can resort to self-normalized IS (21).

Besides being a variance reduction technique, IS can also be seen as a two step sampling procedure for producing samples (approximately) drawn from p, out of samples originally drawn from q. The technique is known as Rubin's SIR mechanism (15) (19) (49) (7, §9.2): Let $\{x_i\}_{i=1}^N$ be iid samples from $q(x)$ and let $\{\tilde{x}^i\}_{i=1}^M$ be M iid samples from $\sum_{i=1}^N \frac{p(x_i)/q(x_i)}{\sum_{i=1}^N p(x_i)/q(x_i)}\delta_{x_i}(\mathrm{d}x)$ (in other words, we draw samples from q, weight each proportionally to $w^u(x_i) = \frac{p(x_i)}{q(x_i)}$, and *resample* iid points from this random discrete distribution). Then $\{\tilde{x}^i\}_{i=1}^M$ become iid. samples from p if $N \to \infty$.

2.1.3 Parametric classifier by minimizing the BCE

From now on we consider the setting where λ, p_1 and p_0 are unknown, and we only have the set $\mathcal{D}$ of labeled samples from p_0 and p_1 (resp. with labels $k = 0, 1$), see §2.1.1. In this context, for classification purposes we should build a parametric function $r_\theta(x)$ that approximates the posterior PDF. The aim of this section is to show that minimizing a BCE loss indeed yields such a suitable approximation. To see this, first recall the definition of the BCE criterion:

$$\mathcal{L}_{\text{BCE}}(\theta) = -\sum_{i=1}^{N_1} \log(r_\theta(x_i^{(1)})) - \sum_{i=1}^{N_0} \log(1 - r_\theta(x_i^{(0)})), \tag{2.10}$$

where $r_\theta(x) \triangleq \Pr_\theta(k = 1|x)$ is the probability under model θ that the label associated to an observation x is 1. Let $h(x, k)$ be the joint distribution over observations and labels:

$$h(x, k) = \underbrace{\frac{N_k}{N_1 + N_0}}_{h(k)} \underbrace{p_k(x)}_{h(x|k)}, x \in \mathbb{R}^d, k = 0, 1. \tag{2.11}$$

Using $r_\theta(x)$, we can build another joint distribution $h_\theta(x, k) = h(x)r_\theta(x)^k(1 - r_\theta(x))^{1-k}$, where $h(x)$ is the x-marginal in (3.3). The BCE is then, up to additive and multiplicative

constants, nothing but an approximation of

$$D_{\mathrm{KL}}(h(x,k)||h_\theta(x,k)) = \mathbb{E}_{h(x,k)}[\log(h(k|x))] - \mathbb{E}_{h(k,x)}\big[\log(\Pr_\theta(k|x))\big], \tag{2.12}$$

where only the last term depends on θ. We indeed retrieve the BCE with the MC approximation:

$$\begin{aligned}\mathbb{E}_{h(k,x)}[\log(\Pr_\theta(k|x))] &\overset{(3.3)}{=} \sum_{k=0,1} \frac{N_k}{N_1+N_0}\mathbb{E}[\log(\Pr_\theta(k|x)] \\ &\approx \frac{1}{N_1+N_0}\Big(\sum_{i=1}^{N_1}\log(r_\theta(x_i^{(1)})) + \sum_{i=1}^{N_0}\log(1-r_\theta(x_i^{(0)}))\Big).\end{aligned}$$

So $\arg\min_\theta D_{\mathrm{KL}}\big(h(x,k)||h_\theta(x,k)\big) \approx \arg\min_\theta \mathcal{L}_{\mathrm{BCE}}(\theta)$.

The interest of this interpretation is that, as is well known, a D_{KL} equals zero when the two distributions are equal almost surely. So, if r_θ represented any arbitrary function, minimizing $D_{\mathrm{KL}}\big(h(x,k)||h_\theta(x,k)\big)$ would ensure that $r_\theta(x)^k(1-r_\theta(x))^{1-k} = h(k|x)$ for all $x \in \mathbb{R}^d$ and $k = 1, 0$, ie that the classifier reaches the target posterior PDF. Of course, in practice, minimizing the BCE does not ensure that this D_{KL} reaches zero. First, we only dispose of a finite number of labeled observations and minimizing an MC approximation of the D_{KL} does not minimize the D_{KL} itself. Next, the parametric family does not contain $h(k|x)$ in general, in which case we can only ever reach a positive minimum of the D_{KL}. Lastly, standard optimization techniques would only guarantee convergence to a positive local minimum of the D_{KL}. Therefore in practice, minimizing the BCE loss only yields an approximates the unknown posterior.

2.1.4 Using a binary classifier for (approximate) Sampling

We now come to the heart of this section. If p_1 is a PDF of interest in an MC sampling setting, and p_0 a suitable easy-to-sample instrumental distribution - be it the proposal in AR, the independent kernel in IMH, or the importance distribution in IS; then the three sampling algorithms involve the PDF ratio $p_1(x)/p_0(x)$, which is unknown when at least one PDF is intractable. As explained in section 2.1.3, a parametric binary classifier trained from a set $\mathcal{D}$ of labeled observations computes an approximation of the unknown posterior distribution. However, remember that (2.6) is equivalent to (2.7); we thus see that classifiers can also be used for approximating *PDF ratios* of interest, which enables us to propose approximate versions of the sampling algorithms based on this classifier-ratio approximation, *and thus to relax the requirement of tractable PDF, but at the cost of approximate sampling.* Of course, the closer p_0 is to p_1, the more efficient the sampling algorithms; however, here p_0 is supposed to be given and our problem is not to adjust p_0 from a given p_1, but rather leverage $\mathcal{D}$ in the case where p_0, p_1 are fixed but unknown PDFs.

Assumptions.

p_1 is the distribution of interest and p_0 a fixed instrumental distribution from which we can propose samples. Ratio $p_1(x)/p_0(x)$ is unknown, but we dispose of the labeled dataset $\mathcal{D}$, and assume that we can train a binary classification model r_θ which minimizes (2.10).

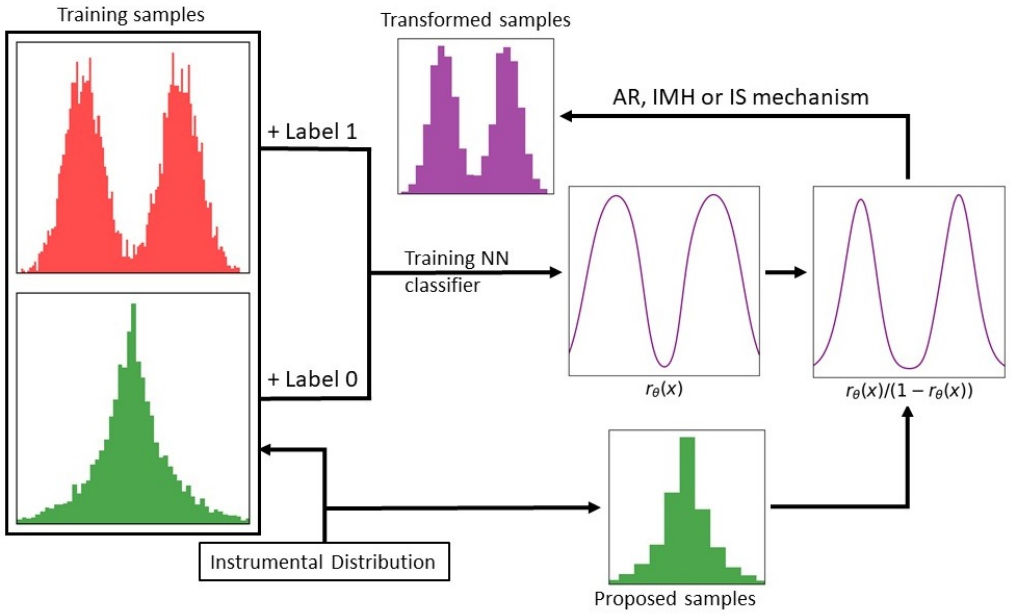


Figure 2.2: Summary of the classifier-based sampling approach

Classifier-based sampling algorithms

A key ingredient for running the algorithms of §2.1.2, is the ratio $p_1(x)/p_0(x)$ which appears in $\alpha_{AR}(x), \alpha_{IMH}(x, x_t)$ and in $w^u(x)$, so following the idea expressed in (2.7), we can however make use of a classifier for approximating the unavailable ratio by making the following substitutions:

$$\alpha_{AR}(x) \leftarrow \frac{1}{\widetilde{C}} \frac{r_\theta(x)}{1 - r_\theta(x)} \text{ where } \widetilde{C} = \max_{y \in \mathcal{D}} \frac{r_\theta(y)}{1 - r_\theta(y)}; \tag{2.13}$$

$$\alpha_{IMH}(x, x_t) \leftarrow \min\left(1, \frac{r_\theta(x)(1 - r_\theta(x_t))}{(1 - r_\theta(x))r_\theta(x_t)}\right); \tag{2.14}$$

$$w^{\mathrm{u}}(x) \leftarrow \frac{r_\theta(x)}{1 - r_\theta(x)}. \tag{2.15}$$

Our procedure is summarized by fig. 2.2: we first train r_θ from labeled samples from p_1 and p_0; we next use ratio $r_\theta(x)/(1 - r_\theta(x))$ as a surrogate of $p_1(x)/p_0(x)$, which enables us to use the AR, IMH or IS procedure, and thus to transform samples from p_0 into (approximate) samples from p_1 via a stochastic operation. *A main advantage of our approach is that a distribution which is only defined by its sampling procedure and has implicit PDF can be used as instrumental* p_0. Indeed our approach does not require evaluating the PDF p_0 neither during the training of the classifier, nor in the proposed sampling procedures.

Illustrating examples

We illustrate our approach (see fig. 2.3) on reference 2D examples in order to illustrate the mechanism of (i) obtaining an approximate of the PDF ratio from samples

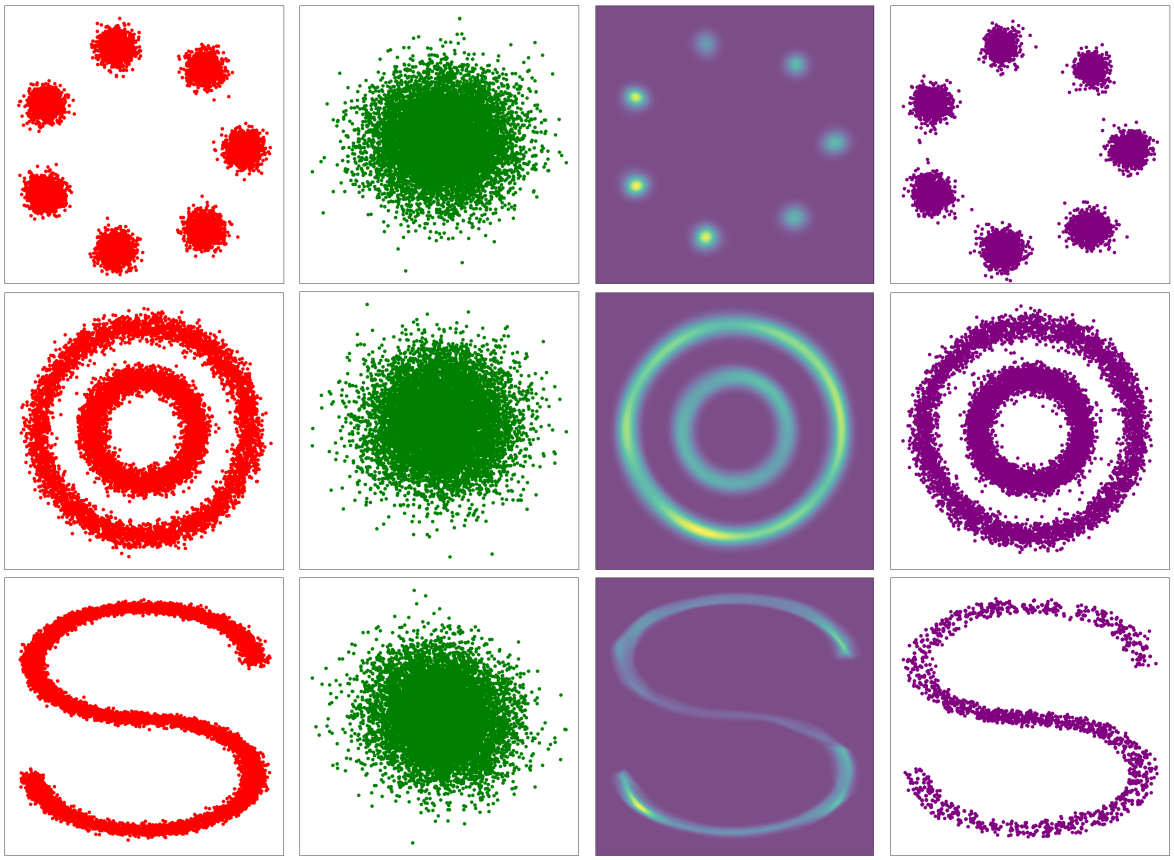


Figure 2.3: Density ratio (middle-right) via classification of samples from p_1 (left) and p_0 (middle-left) - approximate samples from p_1 (right) obtained via a ratio based algorithm: AR (top), IMH (middle), IS (bottom)

using a feed-forward NN (4) with 3 hidden layers, 32 hidden units per layers and SiLU activation function that outputs $\text{logit}(r_\theta(x))$ trained according to the BCE criterion; and (ii) sampling from the target distribution via that PDF ratio using the AR, IMH or IS samplers. The instrumental p_0 was set to be a Gaussian with mean and covariance estimated from the samples from p_1 (even though it can be computed, PDF p_0 was not used during the procedure).

Probabilistic modeling

So far, we have presented our work as a technique to perform approximate MC sampling; let us now revisit it under the scope of probabilistic modeling. If we rewrite p_1 as $p_1(x) = \frac{p_0(x)(p_1(x)/p_0(x))}{\int p_0(z)(p_1(z)/p_0(z))\mathrm{d}z}$, then using (2.7) amounts to building an approximation p_θ of p_1:

$$p_\theta(x) = \frac{p_0(x)(r_\theta(x)/(1-r_\theta(x)))}{\int p_0(z)(r_\theta(z)/(1-r_\theta(z)))\mathrm{d}z}. \tag{2.16}$$

Our procedure consists in applying the AR, IMH or IS samplers to p_θ with proposal p_0 (at least up to the approximation of constant C in the AR case, see (2.13)). This construction corresponds to a specific energy-based model (60)(9)(31) with an energy

function given by:

$$E_\theta(x) = -\log(p_0(x)) - \text{logit}(r_\theta(x)). \tag{2.17}$$

Model p_θ inherits the advantages of this energy structure: (i) it can be trained without evaluating the gradient of the numerator of (2.16) nor of the intractable normalizing constant; (ii) it is structurally compatible with the AR, IMH or IS samplers as sampling from p_θ with instrumental p_0 is equivalent to applying the approximate sampling presented in Section 2.1.4. In Fig. 2.4, we display such an approximation of the distribution of an image. $\text{logit}(r_\theta(x))$ defined via an NN function with 3 hidden layers of size 512 (SiLU activation function), trained according to the BCE criterion, produces a ratio based energy model able to capture details of the target distribution. Samples can effortlessly be obtained via any of the three samplers, with target p_θ and instrumental p_0. Note that computing the unormalized PDF in (2.16) (displayed in the middle-right in Fig. 2.4) indeed requires evaluating PDF p_0 but, again, it is not required for sampling.

Figure 2.4: Classifier based energy model: unormalized PDF (middle-right) and samples (right); obtained from samples (middle-left) from a grayscale image (left) 2D distribution.

2.1.5 Conclusion

The classical AR, IMH and IS samplers require that both the target p_1 and the easy-to-sample instrumental p_0 are known functions. In practice however, both functions may be either unknown (for p_1) or untractable (for p_0). We observed that these samplers use p_1 and p_0 only via their ratio $\frac{p_1}{p_0}$ which, in turn, can be approximated by a classifier. We thus showed that one can still approximately sample from p_1 using AR, IMH or IS in the situation where we can not evaluate the functions p_1 and/or p_0, provided that we dispose of a classifier function, which can be obtained from a set of labeled samples from both distributions. The advantages or our approach are twofold: (i) it is completely PDF-free as compared to standard approaches (neither p_1 nor p_0 needs to be known explicitly); (ii) training reduces to a parametric classification task. From a probabilistic modeling perspective, our approximate samplers coincide with the original ones when applied to some specific energy based approximation of target p_1 which,

thanks to its specific structure, can both be trained easily via standard classification, and is structurally compatible with the AR, IMH or IS sampling techniques.

2.2 Application of the method to the posterior

In the context of the LTER, we can easily apply the proposed methodology to sample from the corresponding approximation of the posterior PDF. Indeed, a simple rewriting of equation (2.2) yields:

$$\frac{p_{X|Y=y}(x)}{p_X(x)} = \frac{p_{Y|X=x}(y)}{p_Y(y)} \approx \frac{r_\theta(x,y)}{1 - r_\theta(x,y)}. \tag{2.18}$$

So the LTER approximation is also an approximation of the posterior-to-prior ratio. So we can perform approximate sampling from the posterior distribution, so $\mathcal{P}$ corresponds to $\mathcal{P}_{X|Y=y}$ with using the prior as instrumental distribution, so $\mathcal{Q}$ is $\mathcal{P}_X$ so long as the prior is easy-to-sample-from (and, as mentioned, not necessarily with tractable PDF).

So we now suppose that (i) the prior distribution $\mathcal{P}_X$ is easy-to-sample from, and (ii) that, using $\mathcal{D}$, we are able to obtain an approximate of the posterior-to-prior PDF based on a classifier r_θ. Following the idea expressed in the section 2.1, we now explicit the classifier-ratio based approximate sampling from the posterior PDF using the AR approach, while the other algorithms are detailed in the appendix of this thesis, section A.3.

Algorithm 2 Classifer based Accept-Reject sampling from the posterior

Require: observed y, $\mathcal{D}$, prior $\mathcal{P}_X$, classifier r_θ
  Compute $\widetilde{C} = \max_{x_i \in \mathcal{D}} \frac{r_\theta(x_i,y)}{1-r_\theta(x_i,y)}$
  while not enough samples are accepted **do**
    Propose $x' \sim \mathcal{P}_X$
    Update $\widetilde{C} = \max(\widetilde{C}, \frac{r_\theta(x',y)}{1-r_\theta(x',y)})$
    if $u \sim \mathcal{U}_{[0,1]} \leq \frac{1}{\widetilde{C}} \frac{r_\theta(x,y)}{1-r_\theta(x,y)}$ **then**
      Accept x'
    end if
  end while

2.3 Connection with energy-based modeling

In the end of the section 2.1, we connected our work to Energy-based modeling and we now elaborate on this connection. Energy-based modeling (1) is a set technique for approximating a target distribution $\mathcal{P}$ using a parameterized probability distribution. The specificity of this estimation technique is that the underlying parameterized probability distribution is not constructed as a generative model but instead, inspired by the principle of Boltzmann distribution in statistical physics, defines the log-probability as the inverse of an *energy* function (hence the term *Energy*-based model). On the one

hand, (deep-) generative models are usually defined by, and parameterized via, their sequential sampling procedure. As such, these models can be expressed as directed graphical models which define a built-in, easy-to-carry step-wise sampling procedure, and the goal of parameter estimation in this context is to adjust the sampling steps so that the underlying probability distribution $\mathcal{P}_\theta$ best matches $\mathcal{P}$. However, as we will see in Chapter 4 of this thesis, which is dedicated to this topic, a generative model does not necessarily admit a closed form expression for its PDF, which also raises practical issues. On the other hand, an EBM is not defined via a sampling procedure but is instead parameterized via its unnormalized PDF directly. Thus, an EBM constructs a parameterized probability distribution $\mathcal{P}_\theta$ via a PDF of the form:

$$p_\theta(x) = \frac{\widetilde{p_\theta}(x)}{\int \widetilde{p_\theta}(x)\mathrm{d}x}. \tag{2.19}$$

The advantage of such a construction is that one can easily use flexible NNs functions for parameterizing the probability distribution, making it suitable for different learning tasks such as generative modeling (17) or classification (24). This probability distribution is parameterized by θ via the unnormalized PDF $\widetilde{p_\theta}(x) \geq 0$ which can be implicitly defined via a function $E_\theta(x)$ which we refer to as energy function (this is detailed in section 2.3.2). Suppose, for instance, that we dispose of a dataset $\mathcal{D} = \{x_i \sim \mathcal{P}\}$ and that we wish to obtain a suitable model θ by maximum likelihood estimation. In order to place our contribution in this specific context, we briefly go over the principle of using EBM, but the literature concerning EBMs is vast and applied in many settings (see, most notably, (41) in the context of image generation, (60) in the context of anomaly detection, (54) in the context of source separation in signal processing, and (53) in the context of scene graph generation).

2.3.1 The issue of the unknown normalizing constant for Maximum Likelihood Estimation

The PDF (2.19) indeed sums to 1 but, for any input x, can only be computed up to its (normalizing) constant denominator since the integral $Z_\theta = \int \widetilde{p_\theta}(x)\mathrm{d}x$ does not admit a closed-form expression in general. This normalizing value, sometimes referred to as the partition function, is a constant with respect to x but it indeed depends on θ and so its intractability poses a real challenge when trying to adjust the parameters of the model. First, for arbitrary $\widetilde{p_\theta}(x)$, no closed-form expression for the value θ which maximizes the (log-) likelihood $p(\mathcal{D}|\theta)$ is available, and so in practice, we rather resort to adjusting θ via a gradient-based optimization. The gradient of the log-likelihood of $\mathcal{D}$ reads:

$$\nabla_\theta \log(p(\mathcal{D}|\theta)) = \sum_{i=1}^{|\mathcal{D}|} \nabla_\theta \log(\widetilde{p_\theta}(x_i)) - |\mathcal{D}|\nabla_\theta \log(Z_\theta). \tag{2.20}$$

This expression for the gradient allows us to understand the effect of a step in a gradient ascent of the log-likelihood as a combination of two actions: (i) increasing the probability

mass $\widetilde{p_\theta}(x)$ in the specific regions of space where the samples from $\mathcal{D}$ lie (the first term), while (ii) globally reducing the probability mass in the support of $\mathcal{P}_\theta$.

However, this expression for the gradient is again unavailable because of the intractable gradient of the log-normalizing constant. It can nonetheless be approximated with MC (59) as:

$$\nabla_\theta \log(Z_\theta)|_{\theta=\theta_t} = \mathbb{E}_{X\sim\mathcal{P}_{\theta_t}}\left[\nabla_\theta \log(\widetilde{p_\theta}(X))|_{\theta=\theta_t}\right] \tag{2.21}$$

$$\approx \frac{1}{M}\sum_{i=1}^{M} \nabla_\theta \log(\widetilde{p_\theta}(x_j^{(t)}))|_{\theta=\theta_t}, \text{ where } x_1^{(t)}, ..., x_M^{(t)} \overset{iid}{\sim} \mathcal{P}_{\theta_t}, \tag{2.22}$$

which yields a gradient update that can again be interpreted as a combination of two actions. The gradient reads in this case:

$$\nabla_\theta \log(p(\mathcal{D}|\theta))|_{\theta=\theta_t} \approx \nabla_\theta \sum_{i=1}^{|\mathcal{D}|} \log(\widetilde{p_\theta}(x_i))|_{\theta=\theta_t} - \frac{|\mathcal{D}|}{M}\sum_{j=1}^{M} \nabla_\theta \log(\widetilde{p_\theta}(x_j^{(t)}))|_{\theta=\theta_t}. \tag{2.23}$$

On the one hand, the first term yields an increase in the probability mass in the regions where lie the observations $\mathcal{D}$. On the other hand, the second term yields a decrease of the probability mass in the region where lie the samples of the current model $\mathcal{P}_{\theta_t}$. Therefore, an efficient and accurate sampling of the model is required during the maximum likelihood training procedure. Of course, since the PDF $p_\theta(x)$ can be evaluated up to a constant for any value of θ, one can sample from the underlying distribution using an MCMC method (possibly even gradient informed). In practice, using the data $\mathcal{D}$ from the target distribution as the starting point of a few steps of MCMC transition kernel yields the efficient contrastive divergence training algorithm for (30) (see also (29) for a practical guide), which was studied and improved upon in subsequent work such as (57) (58) (44). In this context, other sampling methods such as IS (38) and SMC sampling (8) have also recently been applied in an attempt to improve the training of an EBM.

2.3.2 Appropriate parameterization of the unnormalized PDF via an Energy function

In order to obtain an EBM which is practical to use and to train, the first step is to parameterize the unnormalized PDF accordingly. We usually parameterize $\widetilde{p_\theta}(x)$ using a positive function $E_\theta(x)$, which is referred to as *Energy* function:

$$\widetilde{p_\theta}(x) \triangleq \exp(-E_\theta(x)). \tag{2.24}$$

Of course, MCMC methods enable sampling from $\mathcal{P}_\theta$ via the unnormalized PDF (or possible via its gradient), but an appropriate choice of the energy function may enable a convenient sampling scheme. This question has, both historically (32) and recently (37)(3), motivated the development of sampler-induced EBM and we now revisit the Binary-Classification Monte Carlo sampling methodology in this context.

The example of Restricted Boltzmann Machines (RBM)

We first illustrate how an appropriate choice of energy function can induce a convenient sampling scheme with the example of the Restricted Boltzmann Machine (RBM) in the context of binary data $x \in \{0,1\}^d$. RBMs, which were introduced in (47) (see (61) and the references therein for a thorough review), are EBMs where the energy function reads:

$$E_\theta(x) = -(a^T x + \sum_{j=1}^{p} \log(1 + e^{(W_j x + b_j)})); \tag{2.25}$$

and where parameter θ comprises $W \in \mathbb{R}^{p \times d}, a \in \mathbb{R}^d, b \in \mathbb{R}^p$. RBM can be understood as a particular instance of Boltzmann Machines (32) with a specific structure which induces a computational convenience that enables sampling from the underlying distribution easily via a Gibbs scheme. Indeed, this specific expression for $E_\theta(x)$ can be obtained as the energy function associated with the x-marginal distribution of a joint EBM with energy function:

$$E_\theta(x, z) = -z^T W x - x^T a - z^T b, \text{ with } z \in \{0,1\}^p. \tag{2.26}$$

It follows that this specific choice of energy function produces an EBM where $p_\theta(x, z), p_\theta(x)$ and $p_\theta(z)$ are all only available up to a normalizing constant and cannot be expressed in closed form as some distribution which we can easily sample from. However, it is very convenient that both conditional distributions $p_\theta(x|z)$ and $p_\theta(z|x)$ can be written as a product of independent univariate Bernoulli distributions:

$$p_\theta(x|z) = \prod_{i=1}^{d} p_\theta(x_j|z) \text{ where } \Pr(x_i = 1|z) = \text{sigmoid}(\mathrm{W_{i,\_}z + a_i}); \tag{2.27}$$

$$p_\theta(z|x) = \prod_{j=1}^{p} p_\theta(z_j|x) \text{ where } \Pr(z_j = 1|x) = \text{sigmoid}(\mathrm{x^T W_{\_,j} + b_j}) \tag{2.28}$$

where $W_{i,\_}, W_{\_,j}$ are respectively the i-th row and j-th column of W. Indeed, the fact that these two conditional distributions are available in closed form enables us to use a Gibb sampling MCMC scheme. Therefore, by sequentially sampling the two conditionals, we obtain samples from the joint distribution $p_\theta(x, z)$ and hence the x values produced are samples from the marginal distribution of interest $p_\theta(x)$.

An instrumental-distribution based energy function

With that regard, the methodology presented in the section "Binary classifier based Monte Carlo sampling" is also relevant in the context of EBM. Indeed, as mentioned in section 2.1.4, the methodology we proposed to perform approximate sampling from the distribution of interest $\mathcal{P}_1$ can also be understood as applying the AR, IMH, or IS sampling principle to obtain samples from $\mathcal{P}_\theta$ an EBM approximation of $\mathcal{P}_1$. This model has PDF which, up to a normalizing constant, reads:

$$\widetilde{p_\theta}(x) = \frac{r_\theta(x)}{1 - r_\theta(x)} q(x); \tag{2.29}$$

so this underlying probability distribution corresponds to an EBM with the energy function $E_\theta(x) = -\text{logit}(r_\theta(x)) - \log(q(x))$. The energy $E_\theta(x)$ and consequently the unnormalized PDF (related to the energy function via (2.24)) can be evaluated so long as $q(x)$, the PDF associated to the instrumental distribution $\mathcal{Q}$, can be computed up to a constant. However, we will explain hereafter that this is not required in the methodology we propose, neither for maximum likelihood estimation of θ, nor for sampling from the corresponding distribution.

The advantage of the specific structure of this energy function is that it enables us to draw samples from the underlying probability distribution with the algorithms of AR, IMH and IS using the instrumental distribution $\mathcal{Q}$. Indeed, because the unnormalized PDF is, up to a constant, the product of an instrumental PDF $q(x)$ and $r_\theta(x)/(1 - r_\theta(x))$, the approximation of the ratio $p_\theta(x)/q(x)$, does not involve neither $q(x)$ nor the normalizing constant Z_θ. The consequence of this is twofold: firstly, the PDF ratio can be evaluated exactly as $r_\theta(x)/(1 - r_\theta(x))$, thus enabling us to apply the classical ratio-based MC sampling methods. Second, as mentioned in the section 2.1, this methodology therefore enables the use of an instrumental distribution which is an implicit generative model since the sampling steps do not involve PDF $q(x)$.

A training algorithm exempt of normalizing constant estimation

Similar EBM constructions were proposed in the literature. Most notably, the Learned-Accept-Reject Sampling (LARS) methodology (3)(51) has notable connections with our approach. It consists in parameterizing an acceptance probability associated with a truncated rejection sampling approach. LARS has been successfully applied in the context of generative modeling as a way to enrich the base distribution (often called prior distribution in this context) in the contexts of Variational AutoEncoders (36) and of Normalizing Flows (NF) (45). However, as we have mentioned before, for a gradient based maximum likelihood estimation, it is required to account for the fact that the model is unnormalized, and though the normalizing constant can be estimated effortlessly in a differentiable way (with respect to model parameters) in the LARS methodology, it does not circumvent the potential shortcomings associated with a noisy or poor approximation of the log-likelihood. Alternative approaches for estimating the parameters of an EBM which indeed circumvent the problem of the unknown normalizing constant have been proposed in the literature. On the one hand, one can most notably refer to the principle of score-matching (34)(33) which, as an alternative to maximum-likelihood estimation, proceeds by fitting the gradient of the log-PDF (referred to as *Score function*) to the gradient of the unknown PDF via minimizing a square distance between the two functions. On the other hand, (26) proposes to estimate the parameters of an arbitrary EBM by discriminating between components within an implicit mixture distribution. In our approach, the classification procedure is not used to estimate the parameters of an arbitrary EBM but is instead used to construct a specific unnormalized model, and the training procedure reduces to training that classifier to distinguish between the samples from the target distribution and the samples from the instrumental distribution. This approach is therefore free of any estimation of the corresponding normalizing constant.

2.4 A simple, ratio-based, relaxation of bijectivity constraint of NFs

In this section, we consider an alternative problem which is more closely related to the context of Chapter 4 and, more precisely, with the task of generative modeling with NFs. In this context, we propose an easy solution to identify and discard samples that are located in artifact bridges when the target distribution has disjoint elements of mass.

NFs are a popular class of generative models with tractable PDF which have been successful applied in many applications, and have also been used as a tool to build elaborate and efficient MC methods. A complete description is presented in this thesis (see section Chapter 4) but we already briefly describe its principle. An NF applies a change of variable T_θ to a base instrumental distribution $\mathcal{Q}$, and parameters θ are estimated such that the resulting probability distribution $\mathcal{P}_\theta$ is close to target $\mathcal{P}$ known either via recorded samples or via its PDF.

In this section, we refer to as disjoint a distribution $\mathcal{P}$ for which the set $\mathcal{S}_\epsilon \triangleq \left\{x \in \mathbb{R}^d | p(x) \geq \epsilon\right\}$ is disjoint for small values of ϵ. As an example, a mixture of Gaussian distributions is mutimodal but it can also be disjoint if the mixture components have small variance and are not located close to one another. On the one hand, when the change of variable T_θ is parameterized via a flexible NN function and (ii) the base distribution is non-disjoint, the corresponding NF is known to produce an accurate approximation of non-disjoint target distributions. On the other hand, if the base distribution is non-disjoint, the corresponding NF, being a continuous transformation of the base distribution, will therefore struggle to efficiently approximate a disjoint distribution. The Normalizing-flow will hence either not be disjoint or only be disjoint for values $\epsilon^{'} \gg \epsilon$, and therefore have significant probability mass in regions where the target distribution has little to no probability mass. We refer to these regions as *bridges* as they connect the different regions of mass. If the NF $\mathcal{P}_\theta$ is an appropriate estimation of the target distribution $\mathcal{P}$ then using it as an instrumental distribution would yield efficient sampling in an MC setting. However, in the context of a disjoint target distribution, an NF can correspond to an accurate estimation in the regions of high probability mass $\mathcal{S}_\epsilon$ in the sense that they appropriately capture the distribution in these regions; but be inaccurate in some regions of low probability mass because of bridges where the NF model has high probability mass. We can therefore summarize this by saying that an NF is approximately proportional to the target distribution in $\mathcal{S}_\epsilon$ for a small value of ϵ.

Previous approaches have been proposed to tackle this issue and remove or prevent the occurrence of bridges in NFs (16)(13)(11)(51) and in Chapter 4 of this thesis, we propose the novel construction of *Discretely Indexed Flows*, a parametric probability distribution which does not suffer from the same structural limitation in the context of disjoint distributions. In the context of NFs, we now propose a simple, yet surprisingly efficient method to remove bridges from an NF model and, more precisely, discard the samples from the model which are located on these bridges. This approach is once again based on the PDF ratio $p(x)/p_\theta(x)$, where p_θ is the PDF associated with the NF model. First, the value of this PDF ratio can, in this context, indicate whether or not a sample $x \in \mathbb{R}^d$ is located on a bridge. On the one hand, if the PDF ratio takes a high value

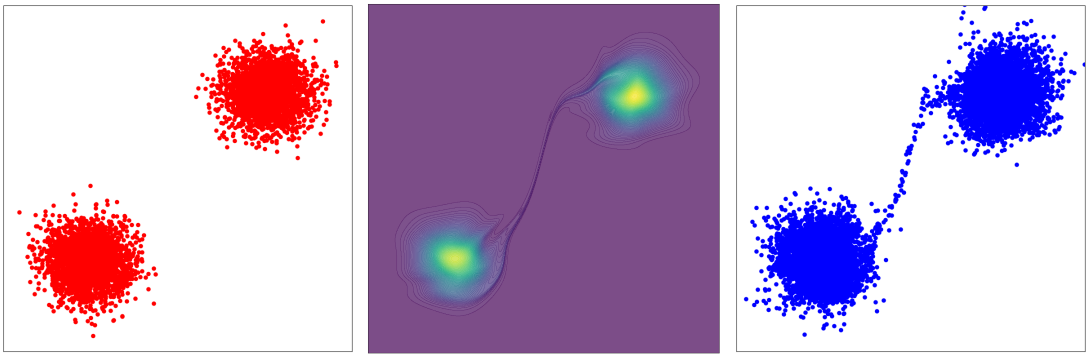


Figure 2.5: Non-disjoint NF (PDF (Middle) and samples (Right)) approximating a disjoint target distribution (Left).

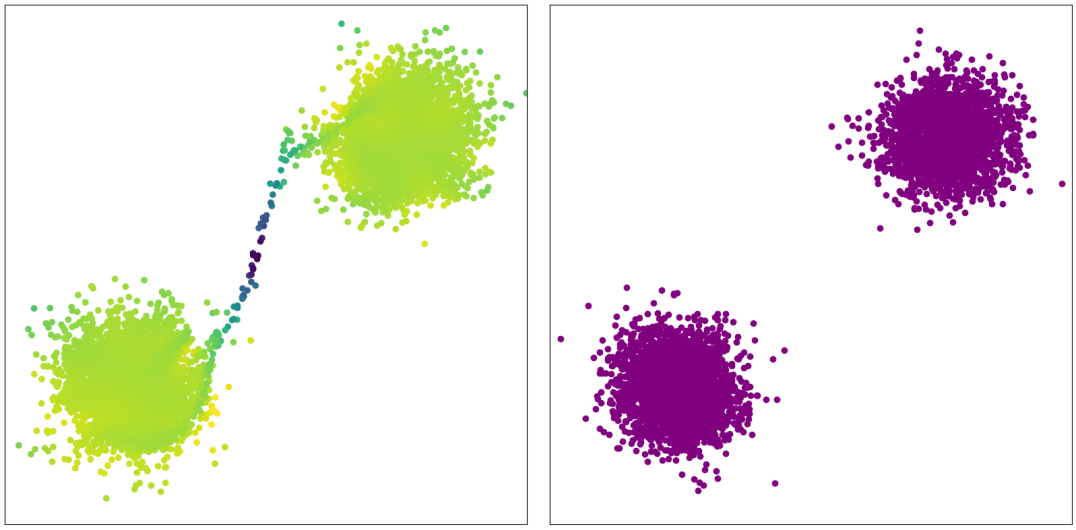


Figure 2.6: Samples from an NF which are located on a bridge can be identified using the PDF ratio (Left) and can be discarded using the proposed resampling strategy (Right).

for a given x, then it means that x is located in a region where $p(x) \geq \epsilon$ hence not on a bridge. On the other hand, if the ratio is small, this means that x is likely located in a region where $p(x)$ is small and $p_\theta(x)$ is not smaller than $p(x)$, which comprises the bridges. So by evaluating the PDF ratio for the samples $x \sim \mathcal{P}_\theta$, the smaller the ratio is, the more likely it is that this sample is on a bridge of the NF, which allows us to identify such regions. Second, we can also easily discard the samples that are in fact located on a bridge with a simple accept-reject step with probability $\alpha(x) = \min(1, \frac{p(x)}{p_\theta(x)})$. This therefore yields a two-step sampling procedure that would only be exact in the case where $p_\theta(x)$ and $p(x)$ are proportional in the support of $\mathcal{P}$ (which is nothing but $\mathcal{S}_\epsilon$ for $\epsilon = 0$). Since in the case of a trained NF, $p_\theta(x)$ is approximately proportional in $\mathcal{S}_\epsilon$ for small ϵ, this procedure is motivated.

NFs are a versatile tool: (i) they can approximate a probability distribution via its PDF (in the context of VI) or from its samples (in the contexts of Generative modeling and DE) and (ii) they provide with tractable PDF and with a straightforward sampling

procedure. As a consequence, if the target distribution is available via its PDF, then the PDF ratio is available which enable to use the proposed technique to remove the bridges from the NFs. Moreover, if the target distribution is available via samples drawn from it, then one can sample from the NF model and obtain an approximation of the PDF ratio using a classifier r_θ which can be obtained via the minimizing the BCE.

2.5 Conclusion

A PDF ratio is a versatile tool in statistical methods. It enables us to compute statistical distances between probability distributions and thus estimate (and minimize) discrepancies between two distributions. In the context of sampling, the PDF ratio is an importance weight which can be used to turn samples from one distribution into samples from another via a stochastic procedure. This principle yields the algorithms of AR, IS, or IMH.

A PDF ratio can be computed if both PDFs in the numerator and denominator can be evaluated (perhaps up to a common constant). However, the value of the PDF ratio can also be obtained from the posterior probability in a binary mixture context. This establishes a connection between PDF ratio evaluation and classification. As a consequence, a possible approach to approximating a ratio is to obtain a classifying parametric model which is trained to distinguish samples from the distribution of the numerator from that of the denominator. This can be achieved, for example, using the BCE criterion computed from recorded samples. In the context of approximate posterior inference, this approach yields the LTER approximation. By using a classifier ratio and the corresponding instrumental distribution, we can build an approximation of a target distribution.

A natural question that arises in this context is that of sampling from the corresponding model, and this topic is the main focus of this chapter. We proposed the "Binary Classification Monte Carlo Sampling" methodology, in which we turned this problem around. We applied the usual sampling algorithms using an instrumental distribution, but where the PDF ratio was replaced by a classifier-based approximation. This yields straightforward and parameter-free sampling schemes, where neither PDFs must be evaluated. As such, the instrumental distribution can be implicit and only requires being sampled from.

We further elaborated on the corresponding model. They are implicit unnormalized energy-based models where the specific choice of energy function provides a straightforward training procedure that circumvents the approximation of the (gradient of the) normalizing constant and is instead a binary classification task. This specific energy function is also structurally compatible with the ratio-based sampling algorithms, yielding the aforementioned sampling procedures of approximate AR, IS, or IMH.

We finally considered classifier ratio in the alternative context of generative modeling, where we proposed to use a classifier ratio for refining NF models. Indeed, NFs are built as a change of variables and preserve the topology of the base distribution. In the case of disjoint distribution, artifact bridges remain to connect the different modes. The PDF ratio can be used to identify and discard the samples from these bridges. This proposed

method can establish a methodological connection between this chapter and "Discretely Indexed Flows", which proposes a model that can be considered an extension to NFs that does not suffer from this topological constraint.

2.6 Perspectives and future work

2.6.1 Bayesian Uncertainty quantification for ratio-based models

Uncertainty quantification is a topic of utmost importance in recent machine learning methods and applications but is not often considered in the context of EBMs. While most methods for uncertainty quantification revolve around bagging and model averaging, Bayesian methods are particularly interesting since studying the posterior predictive distribution enables to unravel differences with regard to the behavior of different models (which is precisely the score point of Chapter 3). When it comes to unnormalized and univariate energy-based models, the PDF of the PPD, which reads:

$$p(x|x_1, ..., x_{|\mathcal{D}|}) = \int \frac{\widetilde{p}_\theta(x)}{Z_\theta} p(\theta|\mathcal{D}) \mathrm{d}\theta; \tag{2.30}$$

is not easy to use in practice. Indeed, even though this distribution can be sampled from, at least in theory, in a two-step procedure with drawing $\theta \sim p(\theta|\mathcal{D})$ and then $x \sim p(x|\theta) = \widetilde{p}_\theta(x)/Z_\theta$; it turns out that the first step can not be conducted easily. As a matter of fact, the PDF $p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)\pi(\theta)$ cannot be computed because of the normalizing constant Z_θ which indeed depends on θ. We recall that a similar issue arises when considering the maximum likelihood parameter estimation problem. So it remains unclear how to compute, approximate or sample from this integral.

As discussed in this chapter, a key advantage of using an approximation based on a ratio and an instrumental distribution is the ability to implement an alternative training procedure. This method avoids the challenge of approximating the normalizing constant, which is a significant burden in maximum likelihood estimation for EBMs. Instead, the training process becomes a binary classification task, distinguishing between samples from the target distribution and samples from the instrumental distribution, labeled as $k = 0$ and $k = 1$, respectively. In Chapter 3, we compare different modeling approaches using the Bayesian PPD). Our methodology is as follows: (i) drawing a graph to show the dependencies between all random variables and deducing a factorization of the joint distribution, (ii) obtaining an expression for the PPD and explaining practical sampling from this distribution; and (iii) analyzing distributions of interest to understand various behaviors in the corresponding inference problem. However, applying this methodology to classifier-ratio based modeling approach presents challenges. Specifically, the first step is not straightforward difficult to draw a Bayesian graph that includes all variables (observations, labels, binary labels, parameters). Therefore, we have not yet analyzed this modeling technique within the Bayesian uncertainty quantification, PPD-based, framework.

Future work could include comparing this classifier-ratio based approach with generative and discriminative modeling techniques, particularly in the context of posterior

learning with the LTER approximation. This comparison could provide insights into the performance of this method in scenarios such as semi-supervised learning, dataset imbalance, and other relevant properties.

2.6.2 Binary Classification based Monte Carlo Sampling: target PDF and implicit instrumental ?

As we have explained, the approximate sampling approaches are completely PDF-free once we dispose of a classifier r_θ. Usually, a classifier r_θ which suitably distinguishes between two probability distributions is obtained via the BCE criterion computed via recorded samples from these two distributions, and the minimization is conducted over the parameters of an otherwise arbitrary function $r_\theta \in [0, 1]$. This function can be obtained by parameterizing:

$$r_\theta(x) = \text{sigmoid}(f_\theta(x)); \tag{2.31}$$

where $f_\theta(x) \in \mathbb{R}$ is an arbitrary function, usually an NN function.

However, usually in the context of MC sampling, the goal is to sample from a distribution which is known via its PDF, perhaps up to a constant. In this specific setting, if the instrumental distribution also has a tractable PDF (perhaps also up to a constant), then the PDF ratio can be computed up to a constant and the usual MC sampling algorithms can be applied. Conversely, it is possible that the instrumental distribution is a suitable candidate for instrumental distribution in the sense that it is selected (or constructed) to be close to the target distribution and easy to sample from, but has an implicit PDF. This situation occurs, for example, when the instrumental distribution is a directed graph which can be sampled from via its latent variables but its PDF is unavailable since one cannot explicitly marginalize the latent variables in the joint PDF. In this setting, the PDF ratio is therefore unavailable, and the classifier-ratio methodology for MC sampling would also prove relevant. We briefly discuss this problem and provide with

We now suppose that we dispose of the target PDF, which in general, is available up to an unknown normalizing constant C such that $p(x) = \frac{\widetilde{p}(x)}{C}$. We also suppose that $\mathcal{Q}$ is a suitable easy-to-sample-from instrumental distribution but with untractable PDF $q(x)$. In this case, the classifying posterior probability can be written (in the case of equally probable a priori classes) as:

$$\Pr(k = 1|x) = \frac{p(x)}{p(x) + q(x)} = \frac{\widetilde{p}(x)}{\widetilde{p}(x) + Cq(x)}. \tag{2.32}$$

In order not to discard the information about target distribution which is contained in $p(x)$, one can parameterize an approximation of (2.32) as:

$$r_\theta(x) = \frac{\widetilde{p}(x)}{\widetilde{p}(x) + \exp(-f_\theta(x))}; \tag{2.33}$$

which is to be opposed to (2.31). This parameterization can easily be motivated: (2.32) is unavailable only because the second term of the denominator, which is a an unnormalized

PDF, is untractable. Instead of approximating the whole posterior probability but rather only the unknown term using a similar construction as an EBM.

By doing so, we retrieve a methodology which is closely related to *Noise-Constrastive Estimation* of an unnormalized model (26), with a similar procedure and a similar training objective. However, the underlying goal is not the same: in the noise-contrastive approach, the classifier-ratio is a tool for estimating an unnormalized (energy-based) model; in our case the approximation of the ratio is obtained via a classifying function which is defined using an unnormalized model.

This procedure obtains the classifier-ratio approximation via the BCE criterion and can thus currently only be applied in the case where we already dispose of recorded samples from the target distribution. However, in usual signal processing and parameter estimation tasks, the goal of MC techniques is often to circumvent the unavailability of samples and/or of a straightforward sampling procedure, usually using an algorithm which is based on the PDF of the target distribution. Therefore, it would be of particular interest to find a way to obtain a classifier-ratio approximation method which does not rely on the BCE criterion (2.10) since it is computed using samples from the target distribution. This question is however, still open and postponed to future work.

Bibliography

[1] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

[2] N. Bartoli and P. del Moral. *Simulation et algorithmes stochastiques*. Cépaduès éditions, 2001.

[3] Matthias Bauer and Andriy Mnih. Resampled priors for variational autoencoders. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 66–75. PMLR, 2019.

[4] George Bebis and Michael Georgiopoulos. Feed-forward neural networks. *Ieee Potentials*, 13(4):27–31, 1994.

[5] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR, 2018.

[6] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, pages 81–88, 2007.

[7] O. Cappé, É. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer-Verlag, 2005.

[8] Davide Carbone, Mengjian Hua, Simon Coste, and Eric Vanden-Eijnden. Efficient Training of Energy-Based Models Using Jarzynski Equality. *arXiv preprint arXiv:2305.19414*, 2023.

[9] Miguel A Carreira-Perpinan and Geoffrey Hinton. On contrastive divergence learning. In *International workshop on artificial intelligence and statistics*, pages 33–40. PMLR, 2005.

[10] Marc Castella, Selwa Rafi, Pierre Comon, and Wojciech Pieczynski. Separation of instantaneous mixtures of a particular set of dependent sources using classical ICA methods. *EURASIP Journal on Advances in Signal Processing*, 2013:1–18, 2013.

[11] Anthony Caterini, Rob Cornish, Dino Sejdinovic, and Arnaud Doucet. Variational inference with continuously-indexed normalizing flows. In *Uncertainty in Artificial Intelligence*, pages 44–53. PMLR, 2021.

[12] Siddhartha Chib and Edward Greenberg. Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49(4):327–335, 1995.

[13] Rob Cornish, Anthony Caterini, George Deligiannidis, and Arnaud Doucet. Relaxing Bijectivity Constraints with Continuously Indexed Normalising Flows. In *Proceedings of the 37th ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 2133–2143. PMLR, 13–18 Jul 2020.

[14] Antonia Creswell and Anil Anthony Bharath. Denoising adversarial autoencoders. *IEEE transactions on neural networks and learning systems*, 30(4):968–984, 2018.

[15] RUBIN DB. Using the SIR algorithm to simulate posterior distributions. In *Bayesian statistics 3. Proceedings of the third Valencia international meeting, 1-5 June 1987*, pages 395–402. Clarendon Press, 1988.

[16] Laurent Dinh, Jascha Sohl-Dickstein, Hugo Larochelle, and Razvan Pascanu. A RAD approach to deep mixture models. *arXiv preprint arXiv:1903.07714*, 2019.

[17] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32, 2019.

[18] Conor Durkan, Iain Murray, and George Papamakarios. On contrastive learning for likelihood-free inference. In *International conference on machine learning*, pages 2771–2781. PMLR, 2020.

[19] Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

[20] James E Gentle. *Random number generation and Monte Carlo methods*, volume 381. Springer, 2003.

[21] John Geweke. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica: Journal of the Econometric Society*, pages 1317–1339, 1989.

[22] Christophe Giraud. *Introduction to high-dimensional statistics*. CRC Press, 2021.

[23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[24] Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*, 2020.

[25] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *J. of Machine Learning Research*, 13(1):723–773, 2012.

[26] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

[27] John Michael Hammersley and David Christopher Handscomb. General principles of the Monte Carlo method. In *Monte Carlo Methods*, pages 50–75. Springer, 1964.

[28] Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free mcmc with amortized approximate ratio estimators. In *International conference on machine learning*, pages 4239–4248. PMLR, 2020.

[29] Geoffrey Hinton. A practical guide to training restricted Boltzmann machines. *Momentum*, 9(1):926, 2010.

[30] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

[31] Geoffrey E Hinton. A practical guide to training restricted Boltzmann machines. In *Neural Networks: Tricks of the Trade: Second Edition*, pages 599–619. Springer, 2012.

[32] Geoffrey E Hinton and Terrence J Sejnowski. Optimal perceptual inference. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, volume 448, pages 448–453. Citeseer, 1983.

[33] Aapo Hyvärinen. Some extensions of score matching. *Computational statistics & data analysis*, 51(5):2499–2512, 2007.

[34] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

[35] H. Kahn and A. W. Marshall. Methods of Reducing Sample Size in Monte Carlo Computations. *J. of the Op. Res. Soc. of Amer.*, 1(5):263–278, 1953.

[36] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[37] John Lawson, George Tucker, Bo Dai, and Rajesh Ranganath. Energy-inspired models: Learning with sampler-induced distributions. *Advances in Neural Information Processing Systems*, 32, 2019.

[38] Meng Liu, Haoran Liu, and Shuiwang Ji. Gradient-Guided Importance Sampling for Learning Binary Energy-Based Models. *arXiv preprint arXiv:2210.05782*, 2022.

[39] A. W. Marshall. The use of multi-stage sampling schemes in Monte Carlo computations. In M. Meyer, editor, *Symposium on Monte Carlo Methods*, pages 123–140, New York, 1956.

[40] Luca Martino, David Luengo, and Joaquín Míguez. *Independent random sampling methods*. Springer, 2018.

[41] Jiquan Ngiam, Zhenghao Chen, Pang W Koh, and Andrew Y Ng. Learning deep energy models. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1105–1112, 2011.

[42] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Tr. on Inf. Theory*, 56(11):5847–5861, 2010.

[43] Art B Owen. Monte Carlo theory, methods and examples, 2013.

[44] Yixuan Qiu, Lingsong Zhang, and Xiao Wang. Unbiased contrastive divergence algorithm for training energy-based latent variable models. In *International Conference on Learning Representations*, 2019.

[45] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[46] Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

[47] David E Rumelhart, James L McClelland, and CORPORATE PDP Research Group. *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations*. MIT press, 1986.

[48] Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of approximate Bayesian computation*. CRC Press, 2018.

[49] Adrian FM Smith and Alan E Gelfand. Bayesian statistics without tears: a sampling–resampling perspective. *The American Statistician*, 46(2):84–88, 1992.

[50] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[51] Vincent Stimper, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Resampling base distributions of normalizing flows. In *International Conference on Artificial Intelligence and Statistics*, pages 4915–4936. PMLR, 2022.

[52] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[53] Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gerard Medioni, and Leonid Sigal. Energy-based learning for scene graph generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13936–13945, 2021.

[54] Yee Whye Teh, Max Welling, Simon Osindero, and Geoffrey E Hinton. Energy-based models for sparse overcomplete representations. *Journal of Machine Learning Research*, 4(Dec):1235–1260, 2003.

[55] Owen Thomas, Ritabrata Dutta, Jukka Corander, Samuel Kaski, and Michael U. Gutmann. Likelihood-free inference by ratio estimation, 2020.

[56] Owen Thomas, Ritabrata Dutta, Jukka Corander, Samuel Kaski, and Michael U Gutmann. Likelihood-free inference by ratio estimation. *Bayesian Analysis*, 17(1):1–31, 2022.

[57] Tijmen Tieleman. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, pages 1064–1071, 2008.

[58] Tijmen Tieleman and Geoffrey Hinton. Using fast weights iuto improve persistent contrastive divergence. In *Proceedings of the 26th annual international conference on machine learning*, pages 1033–1040, 2009.

[59] Laurent Younes. On the convergence of Markovian stochastic algorithms with rapidly decreasing ergodicity rates. *Stochastics: An International Journal of Probability and Stochastic Processes*, 65(3-4):177–228, 1999.

[60] Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. Deep structured energy based models for anomaly detection. In *International conference on machine learning*, pages 1100–1109. PMLR, 2016.

[61] Nan Zhang, Shifei Ding, Jian Zhang, and Yu Xue. An overview on restricted Boltzmann machines. *Neurocomputing*, 275:1186–1199, 2018.

Chapter 3

Generative vs. discriminative Bayesian Posterior learning

Learning a parametric model from a given dataset indeed enables us to capture intrinsic dependencies between random variables via a parametric conditional probability distribution and, in turn, predict the value of a label variable given the observed variables. In this context, an approach for uncertainty quantification is that of Bayesian statistical learning, where model parameters are treated as random variables and marginalized out. This results in an uncertainty-aware inference via the so-called posterior predictive distribution.

When tackling classification and regression problems, this distribution is, most of the time, explicited under a specific construction described as discriminative. However, the discriminative method is not the only way to approximate a posterior probability distribution using a conditional model, and indeed generative modeling is another possible approach. These two constructions differ in their parameterization: the former parameterizes a distribution over label given observation, while the latter does the opposite. It happens that this difference in construction has many interesting consequences with regard to the posterior predictive distribution and the model behavior in the corresponding inference problem.

In this work, we thus undertake a comparative analysis of generative vs. discriminative approaches under the lens of uncertainty quantification via the posterior predictive distribution. Our objective is to compare the ability of both approaches to leverage information from various sources. We assess the role of a prior distribution, explicit in the generative case and implicit in the discriminative case, leading to a discussion on the sensitivity of discriminative models to imbalanced datasets. We next thoroughly examine the role played by the observed variables in the inference, and discuss whether each approach is compatible (or not) with semi-supervised learning. We also provide practical insights and examine how the modeling choice impacts sampling from the posterior predictive distribution. With regard to this, we propose a general sampling scheme enabling supervised learning for both approaches, as well as semi-supervised learning when compatible with the considered modeling approach. Throughout this chapter, we illustrate our arguments and conclusions using the example of affine regression and validate our comparative analysis through classification simulations using neural network-based

models[1].

3.1 Introduction

The statistical learning tasks of classification and regression (85)(39) are paramount in many scientific fields and have gained increasing interest in the big-data and machine learning era. Elaborate methods and tools (5)(9) enable to leverage flexible parametric models in order to capture intrinsic dependencies between related variables and, in turn, predict the value of a variable of interest given the observed values of the others. Many statistical learning methods, both historical (1)(6)(61) and recent (62)(53)(62), can be understood as building an approximate of the conditional probability distribution of a variable of interest (which we refer to as label) given the value of an observed variable. This is usually achieved by considering a parameterized model and adjusting the parameters by minimizing a loss function computed on a dataset comprised of recorded couples of observations and labels.

However, when predicting a label of interest from an observation using a parametric model, committing to a unique value can lead to high imprecision as there are two sources of uncertainty that one needs to account for (19)(48)(44). The *aleatoric* source of uncertainty results from the stochastic nature of the Data Generating Process (DGP) which, as we assume, generates the observation from the label. Moreover, when the DGP is unknown (or at least with an untractable probability density function (PDF)), using an approximate parametric model induces additional uncertainty about the predicted label, which is referred to as *epistemic* (42)(84). This source of uncertainty includes the possible mismatch between the target and the considered modeling parametric family (which is sometimes referred to as model mispecification), as well as the uncertainty about the model parameters resulting from their inference from a finite dataset.

This leads us to the concept of *uncertainty quantification* (UQ), which aims at computing or estimating confidence or credible intervals associated with a prediction. This problem has become of utmost importance in recent years (in particular for applications where providing a measure of confidence is critical), but remains challenging, especially when using neural network (NN)-based models (28)(43). Different methods for UQ have been proposed, including ensemble methods (20), Jackknife/bootstrap methods (23)(2), Laplace approximation methods (77) or Bayesian modeling methods (32)(47). Among them, Bayesian modeling methods have perhaps emerged as the most promising ones (32). They consist of treating the model parameters as random, and these variables are marginalized out to obtain the predicted law of the label given the observation as well as the dataset. This yields the so-called *posterior predictive distribution* (PPD) which, from now on, is the distribution of interest.

Bayesian methods for UQ are now prevalent in many applications and provide a unified framework for many well established techniques which somehow can be related to the task of sampling from the PPD. However, in the literature, the PPD is most often derived with regards to a specific construction that corresponds to a discriminative mod-

[1]We provide all reproducible code and experiments in the Github repository at github.com/ElouanARGOUARCH/Generative_Discriminative_Uncertainty_Quantification.

eling technique. In this discriminative context, the PPD is easy to use as sampling from the corresponding distribution reduces to (i) sampling parameters from their posterior distribution and (ii) evaluating or sampling from the corresponding model. In practice, Bayesian modeling techniques focus on facilitating the first step, which can be particularly challenging. They include building conjugate models for which, by construction, the model parameter posterior distribution is easy to sample from, see e.g. (75)(37). In the case of more elaborate NN-based models (in which case we often refer to Bayesian neural networks (BNN) (59)(63)), prior conjugacy can be leveraged to some extent in methods such as Bayesian last layer (25), but the PPD can no longer be sampled from straightforwardly. Fortunately, Bayesian methods still enable sampling from this distribution (63), be it via MCMC methods (12)(87)(46) or variational inference (35)(29)(80). Finally, methods related to posterior Bootstrap sampling (65)(58)(27)(64)(26) indeed provide theoretically grounded non-parametric approximate sampling from the posterior distribution over model parameters.

However, parametric posterior learning can be addressed using two main approaches. On the one hand, the prevalent *discriminative* modeling consists in parameterizing a distribution over the label given the observation, see e.g. (51)(50)(36). On the other hand, *generative* modeling also indirectly enables to approximate a posterior distribution using a conditional model, via a parameterized distribution over the observed variable given the label, see (76)(71)(72)(60) for some applications. In many learning tasks, the observation is a high-dimensional random variable (RV) when compared to the usually low-dimensional label. Therefore, discriminative models are much more convenient to work with as compared to generative models, as the latter involve modeling a high-dimensional conditional distribution.

Recently, new developments in generative probabilistic modeling (33)(70)(81) now enable to capture intrinsic distributions, and have paved the way towards a renewed interest in the generative modeling approach (69)(45)(90). Some properties and behaviors of both models have previously been compared (see e.g. (66)(83)(62, §9.4)(24)(57)(89)). Yet a comparative analysis of both modeling approaches under the lens of UQ has not been conducted, even though the generative approach induces a different structure of the PPD, which yields different behaviors of the corresponding model in the inference problem. Finally, our aim is to compare the generative and discriminative approaches under the scope of Bayesian UQ via the PPD, and in particular we address the ability of both approaches to leverage information from various sources.

This chapter is organized as follows. In section 3.2 we first provide a precise description of both generative and discriminative constructions. Next in section 3.3, by analyzing the PPD, we explain the different behaviors of the two modeling approaches in an epistemic uncertainty-aware inference. More specifically, in section 3.3.4, we focus our attention on the role of a specific distribution, which can be understood as a prior distribution associated with the PPD, and analyze the ability of each approach to infer using prior information. By doing so, we give clues as to why discriminative models can suffer from imbalanced datasets while generative ones do not, and we confirm this analysis via both illustrative and quantitative simulations. In order to sample from the PPD, especially in the generative case, we provide in section 3.3.5 a general sampling algorithm which is based on a Gibbs scheme and which can easily be applied to both

approaches. Finally, in section 3.4, we specifically discuss the dependency of model parameters to observations, and conclude on the compatibility of each approach with the task of semi-supervised learning, which aims at inferring the model parameters from both labeled and unlabeled datasets. We propose to leverage the corresponding Gibbs sampling scheme to perform Bayesian semi-supervised learning in the generative case. We finally perform simulations in the context of image classification in order to illustrate the arguments in different learning scenarios.

3.2 Supervised learning: Context and objective

Let (X, Y) be a couple of rv related via a DGP $\mathcal{P}_{Y|X}$ which describes the probability distribution of Y given X. The task of prediction consists in retrieving information about an unknown x_0 which (is assumed to have) generated an observed y_0 via the DGP: $y_0 \sim \mathcal{P}_{Y|X}(Y|X = x_0)$. In this chapter, we use the specific nomenclature of a classification problem and we denote Y as *observation* (from the DGP) and, X as the corresponding *label* (though it is not necessarily categorical). The Bayes formula tells us that, once the value y_0 is observed, the information about x_0 is encapsulated in the *posterior* distribution $\mathcal{P}_{X|Y}$ with PDF given by:

$$p_{X|Y}(x_0|y_0) = \frac{p_{Y|X}(y_0|x_0)\pi_{X_0}(x_0)}{p_{Y_0}(y_0)}, \text{where } p_{Y_0}(y_0) = \int p_{Y|X}(y_0|x)\pi_{X_0}(x)\mathrm{d}x, \tag{3.1}$$

where $p_{Y|X}(.|.)$ is the conditional PDF associated with the DGP and $\pi_{X_0}(.)$ is the PDF associated with the distribution which describes our *prior* knowledge about x_0 (we denote prior distributions with letter Π and their PDFs with π). This posterior distribution can be used to obtain pointwise Bayes predictors by minimizing the expectation of a well-designed loss function l (52): $x_0^* = \arg\min_{x_0} \mathbb{E}_{X|Y=y_0}\big[l(X, x_0)\big]$; but in essence, the posterior distribution describes our inability to commit to a singular value of x_0. This source of uncertainty is induced by the random nature of the DGP and is referred to as *aleatoric*.

If both of these PDFs can be evaluated, then (3.1) can be computed at least up to the constant denominator $p_{Y_0}(y_0) = \int p_{Y|X}(y_0|x)\pi_{X_0}(x)\mathrm{d}x$. In many situations however, the PDF associated with the DGP is intractable and consequently (3.1) cannot be evaluated, not even up to a constant. This situation occurs either when (i) we only dispose of a dataset $\mathcal{D}$ (defined in the next paragraph) generated from and the DGP (and so its PDF) is otherwise simply unknown; or (ii) when the DGP is only available via its stochastic simulation procedure which enables obtaining (and augmentin (22)(56)) a dataset $\mathcal{D}$ but its PDF is implicit (15); historical approaches in this setting include the Approximate Bayesian Computation (ABC) methods (16). We consider the first setting and suppose that we dispose of $\mathcal{D}$ generated from the DGP but that we no longer have access to the random sampling mechanism of the DGP making ABC unfeasible. A possible approach to cope with this shortcoming is to resort to an approximation of the intractable posterior using a conditional probability distribution $\mathcal{P}_\theta$ where parameter θ is inferred using observed couples $\mathcal{D} \triangleq \{(x_i, y_i)|x_i \sim \mathcal{P}_X^{\mathcal{D}}, y_i \sim \mathcal{P}_{Y|X}(Y|X = x_i)\}_{i=1}^{|\mathcal{D}|}$. We denote $\mathcal{P}_X^{\mathcal{D}}$ the probability distribution which effectively generated the values in dataset

$\mathcal{D}$ but we stress here that this probability distribution is not necessarily the same as the prior distribution Π_{X_0} (this particular point and its consequences are discussed in details in section 3.3.4). This general formulation includes the usual tasks of statistical parametric learning: we talk about regression (resp. classification) when X is continuous (resp. categorical).

Since a unique parameter estimate of θ (such as Maximum Likelihood Estimates (MLE) or Maximum A Posteriori (MAP) estimates) can be stained with high imprecision if $\mathcal{D}$ is not representative enough of the DGP, we rather consider θ to be a hidden rv, assume a prior knowledge described by distribution Π_Θ, and approximate (3.1) with the ppd (31):

$$p(x_0|y_0,\mathcal{D}) = \int p(x_0,\theta|y_0,\mathcal{D})\mathrm{d}\theta. \tag{3.2}$$

This PDF indeed accounts for the *epistemic* uncertainty which is the uncertainty about the unknown parameter θ (induced by the finite number of recorded samples in $\mathcal{D}$) propagated to x_0 predicted by this model. The ppd (3.2) is computed by integrating out θ in the joint PDF:

$$p(x_0,\theta|y_0,\mathcal{D}) = p(x_0|y_0,\theta)p(\theta|y_0,\mathcal{D}), \tag{3.3}$$

and so $p(x_0|y_0,\mathcal{D}) = \mathbb{E}_\theta[p(x_0|y_0,\theta)|y_0,\mathcal{D}]$ which explains the denomination: it is the average of (posterior) *predictions* $p(x_0|y_0,\theta)$ over probable models under the *posterior* $p(\theta|y_0,\mathcal{D})$. Ultimately, computing the posterior PDF (3.2), or sampling from the distribution if exact computation of the expectation by integration is unfeasible, would enable identifying the probable (aleatoric) values of x_0 that might have generated y_0 via the DGP, while accounting for the modeling (epistemic) uncertainty.

Illustrating running example

Throughout this chapter, we propose to illustrate arguments and conclusions on the continued example of affine regression. Though it serves as an illustrative example, this specific application to affine modeling is, in itself, relevant since affine relationship between variables of interest are most frequent in many science fields and are in many cases, the first considered dependency hypothesis. Let $(X,Y) \in \mathbb{R}\times\mathbb{R}$ be two real-valued rv (assumed to be) related via an unknown DGP. For illustration purposes, we consider a toy setting where the DGP is of the form:

$$Y = \alpha_1 X + \alpha_0 + \sigma_{Y|X}\epsilon \iff p_{Y|X}(y|x) = \mathcal{N}(y;\alpha_1 x+\alpha_0,\sigma^2_{Y|X}). \tag{3.4}$$

We dispose of recorded data $\mathcal{D}$ produced by the DGP, and where the distribution of the x values is $\mathcal{P}^{\mathcal{D}}_X$. The prior knowledge on x_0 is given by the distribution $\pi_{X_0}(x_0) = \mathcal{N}(x_0;\mu_{X_0},\sigma^2_{X_0})$. We chose the prior and the DGP to be conjugated such that the posterior distribution reads:

$$p_{X|Y}(x_0|y_0) = \mathcal{N}(x_0;(\frac{\alpha_1^2}{\sigma^2_{Y|X}}+\frac{1}{\sigma^2_{X_0}})^{-1}(\frac{\alpha_1(y_0-\alpha_0)}{\sigma^2_{Y|X}}+\frac{\mu_{X_0}}{\sigma^2_{X_0}}),(\frac{\alpha_1^2}{\sigma^2_{Y|X}}+\frac{1}{\sigma^2_{X_0}})^{-1}); \tag{3.5}$$

and so this posterior distribution can be used to assess the quality of the inference when comparing the ppd $p(x_0|y_0,\mathcal{D})$ to it; but we otherwise suppose the DGP unavailable. An example of this setting is illustrated in figure 3.1.

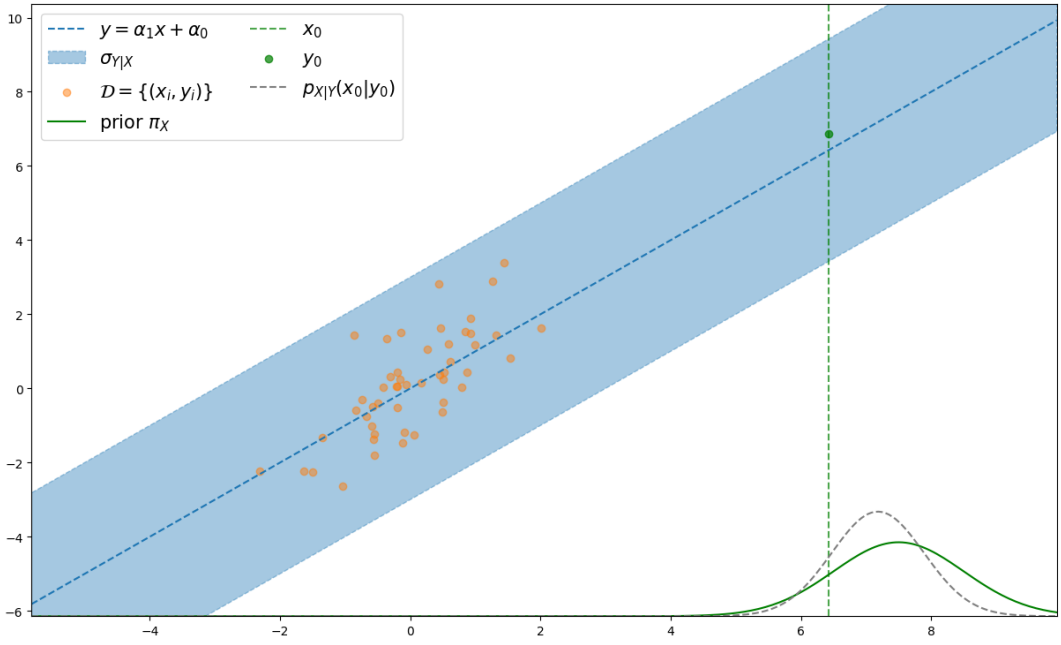


Figure 3.1: Supervised learning setting

3.2.1 Generative versus Discriminative modeling

In the previous section we explained the general principle of modeling the posterior PDF (3.1), and we emphasized on the role of the ppd (3.2), which accounts for the epistemic modeling uncertainty. However, we have not explained precisely yet how the modeling is carried out. In fact, using a parametric conditional probability distribution $\mathcal{P}_\theta$, we can either model the unknown DGP with $\mathcal{P}_\theta(Y|X)$ and deduce the corresponding posterior via the Bayes formula, or model the posterior directly with $\mathcal{P}_\theta(X|Y)$. In the literature, the first approach is classically referred to as *Generative* (since it models the data *generating* process), while the second one is called *Discriminative* (since it makes sense in particular in the classification setting, where the model directly computes the label probabilities which enable to *discriminate* samples via their respective classes). The first approach is called generative modeling but in the (deep) Machine Learning literature, generative modeling (8) can also refer to the task building a parametric probability distribution which is *generative* in the sense that it can be sampled from easily and is designed to resemble a probability which produced recorded data. In this chapter and unless stated otherwise, generative modeling refers to the approach which consists in building an approximate of the posterior distribution of interest (3.1) via modeling the unknown likelihood. These two approaches differ in their philosophy: the first one models only what is unknown, i.e. the generative process, while the second one directly models the function of interest, i.e. the posterior PDF. Figure 3.2 provides with an illustration of the difference between the two approaches.

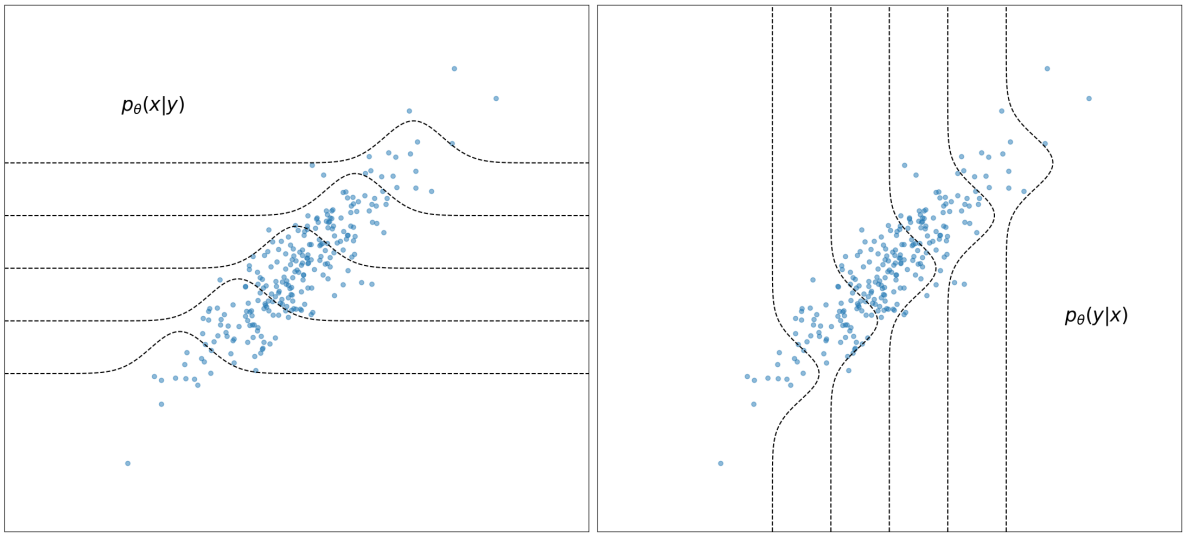


Figure 3.2: Illustration of the difference between generative and discriminative modeling approaches (right)

These approaches yield two different equations:

$$p(x_0|y_0,\theta) \overset{G}{=} \frac{p_\theta(y_0|x_0)\pi_{X_0}(x_0)}{\int p_\theta(y_0|x)\pi_{X_0}(x)\mathrm{d}x} \tag{3.6}$$

$$p(x_0|y_0,\theta) \overset{D}{=} p_\theta(x_0|y_0), \tag{3.7}$$

in which superscripts G and D respectively stand for generative and discriminative; this notation will be used throughout the rest of this chapter. In both equations, p_θ, be it $p_\theta(y_0|x_0)$ in the generative case or $p_\theta(x_0|y_0)$ in the discriminative case, is the conditional PDF associated with $\mathcal{P}_\theta$, and is what is effectively computed with model associated with parameter θ.

Figure 3.3 displays a graphical representation of both models and explains how the modeling choice affects the dependence between all the rv. We build upon these two figures by writing the full joint PDF of all rv. This comparison of graphical models

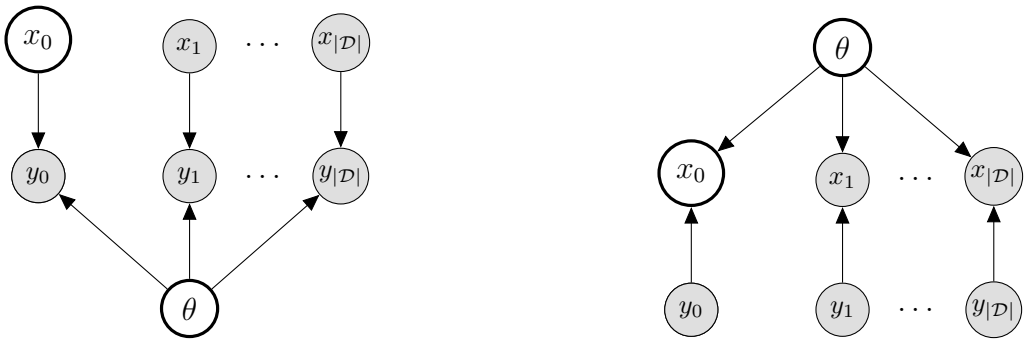


Figure 3.3: Graphical models compared: Generative (left) versus Discriminative (right). The grey (resp. white) nodes are observed (resp. latent) variables.

allow us to deduce the joint distribution of all the rv of interest:

$$p(x_0, y_0, \theta, \mathcal{D}) \overset{G}{=} \pi_\Theta(\theta)\pi_{X_0}(x_0)p_\theta(y_0|x_0) \prod_{(x_i,y_i)\in\mathcal{D}} p_X^{\mathcal{D}}(x_i)p_\theta(y_i|x_i), \tag{3.8}$$

$$p(x_0, y_0, \theta, \mathcal{D}) \overset{D}{=} \pi_\Theta(\theta)p_{Y_0}(y_0)p_\theta(x_0|y_0) \prod_{(x_i,y_i)\in\mathcal{D}} p_Y^{\mathcal{D}}(y_i)p_\theta(x_i|y_i). \tag{3.9}$$

The difference between generative and discriminative modeling has already been established in the literature. In the context of classification, (66) compares the Naive Bayes Classifier (which is a generative model) and logistic regression (which is discriminative) in term of (asymptotic) classification error and conclude in favor of the generative approach when working with a small amount of training data.

Illustrative running example

We now illustrate the difference between both modeling approaches by performing homoskedastic affine regression with unknown variance Gaussian Noise where model parameters are the coefficients of the affine transform as well as the variance of the unknown noise so $\theta = \{\beta_1, \beta_0, \sigma^2\}$. Both the DGP (3.4) and the posterior (3.5) belong to the considered parametric family of conditional probability, so both modeling approaches correspond to a well-specified inference problem. We first describe the generative approach $y_0 \overset{G}{=} \beta_1 x_0 + \beta_0 + \sigma\epsilon, \epsilon \sim \mathcal{N}(0,1) \iff p_\theta(y_0|x_0) \overset{G}{=} \mathcal{N}(y_0; \beta_1 x_0 + \beta_0, \sigma^2)$. Once again, thanks to the conjugacy of the prior $\pi_{X_0}(x_0)$ and the previous $p_\theta(y_0|x_0)$, the posterior PDF $p(x_0|y_0, \theta)$ admits a closed-form expression:

$$p(x_0|y_0, \theta) \overset{G}{=} \mathcal{N}(x_0; (\frac{\beta_1^2}{\sigma^2} + \frac{1}{\sigma_{X_0}^2})^{-1}(\frac{\beta_1(y_0 - \beta_0)}{\sigma^2} + \frac{\mu_{X_0}}{\sigma_{X_0}^2}), (\frac{\beta_1^2}{\sigma^2} + \frac{1}{\sigma_{X_0}^2})^{-1}). \tag{3.10}$$

We now describe the discriminative modeling approach with the same parameterized $\mathcal{P}_\theta$:

$$x_0 \overset{D}{=} \beta_1 y_0 + \beta_0 + \sigma\epsilon, \epsilon \sim \mathcal{N}(0,1) \iff p_\theta(x_0|y_0) \overset{D}{=} \mathcal{N}(x_0; \beta_1 y_0 + \beta_0, \sigma^2). \tag{3.11}$$

3.2.2 Handling multiple observations

The bayesian philosophy behind equation (3.1) is to assume (i) prior knowledge on x_0 (before observation) in the form of a prior Π_{X_0} and (ii) an observation model. This observation model is assumed to produce y_0 from x_0 but it might not necessarily be true in practice. In this case, we say that the observation model is mispecified w.r.t. the observation. In our case, we considered the observation model to be the DGP $\mathcal{P}_{Y|X}$, so (3.1) is a well specified Bayesian setting. Then, after observing (one or) several observations: $y_{0,1}, ..., y_{0,N_0} \overset{iid}{\sim} \mathcal{P}_{Y|X}(Y|X = x_0)$. We deduce the posterior PDF:

$$p_{X|Y_{0,1},...,Y_{0,N_0}}(x_0|y_{0,1}, ..., y_{0,N_0}) = \frac{\pi_{X_0}(x_0)\prod_{n=1}^{N_0} p_{Y|X}(y_{0,n}|x_0)}{\int \pi_{X_0}(x_0)\prod_{n=1}^{N_0} p_{Y|X}(y_n|x_0)\mathrm{d}y_1...\mathrm{d}y_{N_0}}. \tag{3.12}$$

We often describe this equation as a form of Bayesian updating: we update the prior knowledge with the observations. In section 3.3.4, we will discuss the role of the prior Π_{X_0} with regard to both modeling approaches; but in this section, we first specifically examine whether or not each approach enables to easily handle multiple observations in the inference of x_0.

Equations (3.6) and (3.7) explain how we can predict the value of x_0 from a unique observed value of y_0 using model θ for respectively the generative and the discriminative approach. In this case, both approaches enable computation of the posterior $p(x_0|y_0, \theta)$ as both equations are tractable (at least up to a constant). By contrast, when we observe not only a single observation but rather a collection of observations from the DGP which originate from the same unknown value of interest, as in (3.12), then the generative approach enables us to handle this situation with a tractable equivalent of (3.6), while the discriminative one does not.

Indeed, under a generative modeling, we can easily rewrite equation (3.6) as:

$$p(x_0|y_{0,1}, ..., y_{0,N_0}, \theta) \overset{G}{=} \frac{\pi_{X_0}(x_0) \prod_{n=1}^{N_0} p_\theta(y_{0,n}|x_0)}{p(y_{0,1}, ..., y_{0,N_0}|\theta)}; \tag{3.13}$$

and this formula can be computed up to its constant denominator (w.r.t. x_0). On the other hand, with a discriminative modeling, equation (3.7) becomes:

$$p(x_0|y_{0,1}, ..., y_{0,N_0}, \theta) \overset{D}{=} \frac{\prod_{n=1}^{N_0} p_{Y_n}(y_{0,n}) p_\theta(x_0|y_{0,n})}{p(x_0|\theta)^{N_0-1} p_{Y_1,...,Y_{N_0}}(y_{0,1}, ..., y_{0,N_0})}. \tag{3.14}$$

However, factor $p(x_0|\theta) \overset{D}{=} \int p_\theta(x_0|y) p_{Y_0}(y) \mathrm{d}y$ is always intractable since $p_{Y_0}(y)$ given by (3.1) is defined implicitely by the unknown DGP. Therefore, (3.14) cannot be evaluated, not even up to a constant, when $N_0 > 1$. Finally, only the generative approach allows to conveniently deal with multiple observations. In order to carry on with the comparison of both approaches, we only consider the case of a unique observation y_0, but, concerning the generative modeling, all the equations still hold with multiple observations.

3.3 Supervised Epistemic Uncertainty via the ppd

We now discuss how the epistemic uncertainty is accounted for in each approach, be it generative or discriminative. To that end we analyze how the modeling choice impacts the ppd and more precisely how it can be sampled from. We proceed in three steps: first we analyze the model posterior distribution $p(\theta|y_0, \mathcal{D})$ (see §3.3.1), we then deduce the joint distribution $p(x_0, \theta|y_0, \mathcal{D}) \overset{(3.3)}{=} p(x_0|y_0, \theta) p(\theta|y_0, \mathcal{D})$ (see §3.3.2) and we finally come to its (other) marginal of interest, i.e. the ppd (3.2) (see §3.3.3).

3.3.1 model posterior: $p(\theta|y_0, \mathcal{D})$ or $p(\theta|\mathcal{D})$?

In this section we look at the posterior distribution over model θ given the observation y_0 and the recorded dataset $\mathcal{D}$. Using Bayes rule, it can be written as:

$$p(\theta|y_0, \mathcal{D}) = \frac{p(y_0|\theta, \cancel{\mathcal{D}})}{p(y_0|\mathcal{D})} p(\theta|\mathcal{D}), \text{ where } p(\theta|\mathcal{D}) \propto \pi_\Theta(\theta) \prod_{(x_i,y_i)\in\mathcal{D}} p(x_i, y_i|\theta). \tag{3.15}$$

It is important to note here that in the previous equation we can cancel out $\mathcal{D}$ since for any variables involved in figure 3.3, we have $p(.|., \theta, \mathcal{D}) = p(.|., \theta)$. By glancing at these two equations, we can already see that the probable values of θ under this posterior correspond to models for which the elements of $\mathcal{D}$, the couples (x_i, y_i), are likely. In this section, we discuss the impact of y_0 on this distribution and conclude on whether or not this observation carries information for inference of θ depending on the modeling approach. To that hand, we start by leveraging equations (3.8) and (3.9) to deduce:

$$p(y_0|\mathcal{D}) \overset{G}{=} \int p(y_0|\theta) p(\theta|\mathcal{D}) \mathrm{d}\theta \text{ where } p(y_0|\theta) \overset{G}{=} \int p_\theta(y_0|x) \pi_{X_0}(x) \mathrm{d}x, \tag{3.16}$$

$$p(y_0|\theta) \overset{D}{=} p(y_0|\mathcal{D}) \overset{D}{=} p_{Y_0}(y_0). \tag{3.17}$$

On the one hand, with a generative approach, $p(y_0|\theta)$ indeed depends on θ, so y_0 indeed carries information for inferring θ since the two rv are not independent. We can moreover analyze how the information is carried by y_0 to a posteriori models. Probable generative models θ under posterior $p(\theta|y_0, \mathcal{D})$ produce, with high probability, the value y_0 for unknown values x distributed under the prior Π_{X_0}. The posterior distribution of models θ therefore effectively depends on y_0. We finally conclude that the role of y_0 in the ppd inference is twofold: (i) in conjonction to the prior Π_{X_0} it indeed carries information to probable (epistemic) models θ and (ii) it carries information to probable (aleatoric) x_0 values via posterior models θ.

On the other hand, under a discriminative approach, factors $p(y_0|\theta)$ and $p(y_0|\mathcal{D})$ reduce to $p_{Y_0}(y_0)$ (see (3.17)) so θ and y_0 are independent rv and finally the posterior over θ reduces to $p(\theta|\mathcal{D})$. Let us analyze why observation y_0 does not carry any information to a posteriori models. The information carried by y_0 to a discriminative model θ is that it should produce, with high probability, unknown values x for y_0. However, this is nothing but saying that $p_\theta(.|Y = y_0)$ is a probability distribution, which we already know by construction of a discriminative model using $\mathcal{P}_\theta$ a conditional probability distribution. So, by contrast with the generative approach, in the discriminative approach, the role of y_0 is solely aleatoric, i.e. to infer x_0 via probable discriminative models which do not depend on y_0.

3.3.2 Joint PDF

We now derive the joint PDF $p(x_0, \theta|y_0, \mathcal{D})$ given by equation (3.3) for both generative and discriminative modeling approaches. In the generative case (3.6), as explained before, the first factor in the joint PDF is $p(x_0|y_0, \theta) \overset{G}{=} p_\theta(y_0|x_0)\pi_{X_0}(x_0)/p(y_0|\theta)$. In general, this expression can only be computed up to a normalizing constant since

$p(y_0|\theta) = \int p_\theta(y_0|x)\pi_{X_0}(x)\mathrm{d}x$ might be intractable. However, this denominator is a constant w.r.t. x_0 but it indeed depends on θ so it must not be treated as a constant in the joint PDF; so, with regard to the joint PDF, the first factor cannot be computed. Moreover, the second factor is $p(\theta|y_0, \mathcal{D}) \propto p(y_0|\theta)p(\theta|\mathcal{D})$, and as we have explained before in the previous section 3.3.1 indeed depends on y_0. For the same reason, $p(y_0|\theta)$ is intractablethe second factor in the joint PDF cannot be computed, not even up to a constant, either. Conveniently, both factors are intractable because of the same factor $p(y_0|\theta)$ which appears in the denominator of the first and in the numerator of the second. So, even though none of the two factors can be computed individually, the intractable terms cancel out by multiplication and the joint PDF can be computed up to a constant (w.r.t. both x_0 and θ) as:

$$p(x_0, \theta|y_0, \mathcal{D}) \overset{G}{=} \frac{\pi_{X_0}(x_0)p_\theta(y_0|x_0)}{p(y_0|\theta)} \frac{p(\theta|\mathcal{D})p(y_0|\theta)}{p(y_0|\mathcal{D})} \overset{G}{\propto} \pi_{X_0}(x_0)p_\theta(y_0|x_0)p(\theta|\mathcal{D}) \tag{3.18}$$

$$\text{where } p(\theta|\mathcal{D}) \overset{G}{\propto} \pi_\Theta(\theta) \prod_{(x_i,y_i)\in\mathcal{D}} p_\theta(y_i|x_i). \tag{3.19}$$

In the discriminative setting, the first factor in the joint PDF (3.3) reads $p(x_0|y_0, \theta) = p_\theta(x_0|y_0)$ (see again equation (3.7)). This quantity is directly computed in a normalized way by model $\mathcal{P}_\theta$. Moreover, as we pointed out in the previous section 3.3.1, the second factor reduces to $p(\theta|\mathcal{D})$ which can be computed up to a constant. So, unlike in the generative case, both factors can be computed and the joint PDF therefore reads:

$$p(x_0, \theta|y_0, \mathcal{D}) \overset{D}{=} p_\theta(x_0|y_0)p(\theta|\mathcal{D}) \tag{3.20}$$

$$\text{where } p(\theta|\mathcal{D}) \overset{D}{\propto} \pi_\Theta(\theta) \prod_{(x_i,y_i)\in\mathcal{D}} p_\theta(x_i|y_i). \tag{3.21}$$

So in both modeling cases, the joint PDF can be computed (at least up to a constant).

3.3.3 The ppd

Recall that the ppd (3.2) is obtained by marginalizing out the rv θ in the joint distribution (3.3). The consequence is twofold: first its PDF is obtained by integrating the joint PDF w.r.t. variable θ; and second, sampling from the joint distribution provides x_0 samples which are distributed under the ppd. In this section, we discuss the first point.

By integrating (3.18) and (3.20) w.r.t. θ, we obtain expressions for the ppd (3.2) in both cases:

$$p(x_0|y_0, \mathcal{D}) \overset{G}{\propto} \pi_{X_0}(x_0) \int p_\theta(y_0|x_0)p(\theta|\mathcal{D})\mathrm{d}\theta \tag{3.22}$$

$$p(x_0|y_0, \mathcal{D}) \overset{D}{=} \int p_\theta(x_0|y_0)p(\theta|\mathcal{D})\mathrm{d}\theta \tag{3.23}$$

In practice, these two equations can only be used when exact computation of the integral is feasible. Nonetheless, they remain relevant as we can analyze them both to grasp a

difference between the two approaches which provides another interpretation of the ppd. The first formula, in the generative case, corresponds to Bayesian inference using the prior and the marginal likelihood $p(y_0|x_0, \mathcal{D}) = \int p_\theta(y_0|x_0)p(\theta|\mathcal{D})\mathrm{d}\theta$; while the second formula, in the discriminative case, corresponds to an averaging of posterior predictions. So, in both cases, the ppd is an averaging of the quantity p_θ (which is what is effectively computed with model θ) w.r.t. $p(\theta|\mathcal{D})$. This is to be contrasted with the original definition of the ppd, defined as the average of predictions w.r.t. $p(\theta|y_0, \mathcal{D})$. These two definitions only coincide in the discriminative case since the model computes directly the prediction (see (3.7)) and the posterior distribution $p(\theta|y_0, \mathcal{D})$ reduces to $p(\theta|\mathcal{D})$ (see section 3.3.1). As a consequence, this means that $\int p(x_0|y_0, \theta)p(\theta|\mathcal{D})\mathrm{d}\theta$ is an exact construction of the ppd only in the discriminative case.

3.3.4 Explicit or Implicit prior

A main difference between the two approaches lies in the role of the marginal over x_0 in the joint PDF $p(x_0, y_0|\mathcal{D})$. This distribution is of particular interest as it can be considered as a prior $p(x_0|\mathcal{D})$ in the ppd $p(x_0|y_0, \mathcal{D})$ which is the object of interest in both modeling approach:

$$p(x_0|y_0, \mathcal{D}) \propto p(x_0|\mathcal{D})p(y_0|x_0, \mathcal{D}). \tag{3.24}$$

We can again leverage both equations (3.8) and (3.9) to deduce:

$$p(x_0|\mathcal{D}) \overset{G}{=} p(x_0|\theta) \overset{G}{=} \pi_{X_0}(x_0); \tag{3.25}$$

$$p(x_0|\mathcal{D}) \overset{D}{=} \int p(x_0|\theta)p(\theta|\mathcal{D})\mathrm{d}\theta \text{ where } p(x_0|\theta) \overset{D}{=} \int p_\theta(x_0|y)p_{Y_0}(y)\mathrm{d}y. \tag{3.26}$$

These two equations allow us to understand that the marginal over x_0 does not play the same role in the generative and discriminative cases. While in the former setting this immutable marginal distribution describes prior knowledge and does not depend on $\mathcal{D}$ (see equation (3.25)); in the latter setting, this marginal distribution is the result of an intricate interaction between the dataset $\mathcal{D}$, the prior distribution Π_{X_0} and the DGP $\mathcal{P}_{Y|X}$. On the one hand, in the generative approach, it corresponds to the prior Π_{X_0} which can be specified according to the problem at hand and, in itself, may provide significant information about the value of interest x_0. In practice, this prior distribution can also play a role of regularization and may as well be understood as a safeguard since it can effectively constrain the prediction to a specific region of the space (88)(82)(74) but more importantly, the prior distribution is often the result of an elicitation effort (78, Chapter 3) which consists in of (i) obtaining prior information and (ii) transcribing this knowledge into a probability distribution. On the other hand with a discriminative approach, this marginal has a very different role. The relevance of equation (3.26) first lies in the fact that it highlights the systematic intractablity of PDF $p(x_0|\mathcal{D})$. Indeed, it can never be computed (even if exact integration was feasible) since it ultimately involves computing the PDF $p_{Y|X}$ in $p_{Y_0}(y) = \int p_{Y|X}(y|x)\pi_{X_0}(x)\mathrm{d}x$, which is unknown by assumption. This intractability does not pose any practical issue since the computation of the $p(x_0|\mathcal{D})$ (3.26) is not required for computing the joint PDF (3.20) (and

consequently not for computing the PDF of (or sampling from) the ppd). But this also means that the discriminative construction does not allow us to leverage any information encapsulated in, or any practical property induced by, a prior distribution during the inference. We now rewrite the expression of equation (3.26) as:

$$p(x_0|\mathcal{D}) \overset{D}{=} \int\int\int p_\theta(x_0|y)p(\theta|\mathcal{D})p_{Y|X}(y|x)\pi_{X_0}(x)\mathrm{d}x\mathrm{d}y\mathrm{d}\theta \overset{D}{=} \int p(x_0|y,\mathcal{D})p_{Y_0}(y)\mathrm{d}y. \tag{3.27}$$

So, this distribution has a PDF $p(x_0|\mathcal{D})$ which indeed depends on (i) $\mathcal{D}$ via the unknown θ and is indirectly related to (ii) the prior Π_{X_0} and (iii) the DGP $\mathcal{P}_{Y|X}$ via $\mathcal{P}_{Y_0}$. As a consequence, this distribution will attribute high probability mass to the x values which have high probability under $p(x_0|y,\mathcal{D})$ for some value $y \sim \mathcal{P}_{Y_0}$. As such, an implicit density estimation mechanism $x_1, ..., x_{|\mathcal{D}|}$ of $\mathcal{D}$ shifts the distribution $p(x_0|\mathcal{D})$ away from Π_{X_0} and towards the regions of high probability under $\mathcal{P}_X^{\mathcal{D}}$. This implicit density estimation mechanism appears clearly in the limiting case where the aleatoric uncertainty increases since we observe that PDF $p(x_0|\mathcal{D})$ becomes $p(x_0|x_1, ..., x_{|\mathcal{D}|})$. Conversely, when the aleatoric uncertainty decreases, this PDF is, under assumptions of identifibility and invertibility, $\pi_{X_0}(x_0)$. We will illustrate this effect in both contexts of regression (using the running example) and classification. As a consequence, the ppd in the discriminative approach indeed does not provide an approximation of (3.1) with prior Π_{X_0}. It instead provides an approximation of:

$$\frac{p_X^{\mathcal{D}}(x_0)p_{Y|X}(y_0|x_0)}{\int p_X^{\mathcal{D}}(x)p_{Y|X}(y_0|x)\mathrm{d}x}. \tag{3.28}$$

Consequently, a mismatch between the prior Π_{X_0} and the distribution $\mathcal{P}_X^{\mathcal{D}}$, which effectively generated the x_i values in $\mathcal{D}$, will result in a mismatch between the target posterior (3.1) and the ppd (3.2). Subsequently, only the regions of space which are well represented by the x_i values in dataset $\mathcal{D}$ will have high probability mass under the marginal $p(x_0|\mathcal{D})$, and hence, under the ppd $p(x_0|y_0,\mathcal{D})$. Though this argument relates $\mathcal{D}$ to the posterior $p(x_0|y_0,\mathcal{D})$ (via the distribution $p(x_0|\mathcal{D})$), we consider that this argument is not related to epistemic uncertainty as (i) the effect does not vanish when the number of recorded observation, i.e. the size of $\mathcal{D}$ increases; and (ii) the same effect can be observed when considering $p(x_0|y_0,\theta^*)$ where θ^* is a unique parameter (such as MLE or MAP) estimate.

Finally, in the discriminative case, it is of particular interest to study the distribution $p(x_0|\mathcal{D})$ as it corresponds to the average prediction over observations y_0 since:

$$p(x_0|\mathcal{D}) \overset{D}{=} \mathbb{E}_{y_0\sim\mathcal{P}_{Y_0}}\big[p(x_0|y_0,\mathcal{D})\big]. \tag{3.29}$$

This, together with the probability mass of $p(x_0|\mathcal{D})$ which favors the regions of x values in $\mathcal{D}$, tells us that a discriminative model will favor the regions which are well represented in the dataset. In a classification task, the dominant labels will be predicted more often than the others, thus explaining that discriminative models indeed suffer from imbalanced dataset. We further emphasize this precise point using the illustrative running example, a classification example provided in supplementary materials (see section 3.3.4), as well as in quantitative simulations in section 3.5.

Illustrative running example

We now leverage the affine regression example to illustrate the effects of the implicit prior on the ppd in the discriminative modeling approach. We first display an empirical approximation of the distribution $p(x_0|\mathcal{D})$. To that end, using equation (3.27), we obtain samples from this distribution via the two step sampling procedure $y \sim \mathcal{P}_{Y_0}$ and $x_0 \sim p(x_0|y, \mathcal{D})$ (the second sampling step is detailed in the next paragraph 3.3.5). Of course in practice, the first sampling step cannot be conducted as sampling from $p_{Y_0}(y) = \int p_{Y|X}(y|x)\pi_{X_0}(x)\mathrm{d}x$ requires sampling from the DGP which we recall is unknown by hypothesis, but in our example, we do resort to this sampling procedure for illustration purposes.

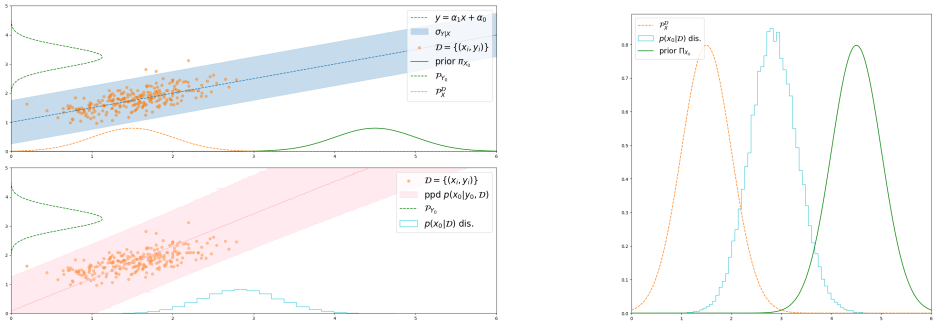


Figure 3.4: Empirical estimate of $p(x_0|\mathcal{D})$ in the discriminative setting: samples are obtained via $y \sim \mathcal{P}_{Y_0}$ (top) followed by $x_0 \sim p(x_0|y, \mathcal{D})$ (bottom). This distribution corresponds to a trade-off between Π_{X_0} and $\mathcal{P}_X^{\mathcal{D}}$ (right)

In figure 3.4, an empirical estimate of $p(x_0|\mathcal{D})$ is obtained via the described two-step sampling procedure (upper-left and lower-left) and is plotted against the prior Π_{X_0} and $\mathcal{P}_X^{\mathcal{D}}$. We see that, in the discriminative setting, $p(x_0|\mathcal{D})$ (which we recall acts as a prior in (3.24)) indeed corresponds to a trade-off between the two distributions and is shifted towards $\mathcal{P}_X^{\mathcal{D}}$ via an implicit density estimation mechanism from the x-values in $\mathcal{D}$. In this example, we can also visualize how the DGP affects the balance between Π_{X_0} and $\mathcal{P}_X^{\mathcal{D}}$ which we now illustrate in the next figure 3.5.

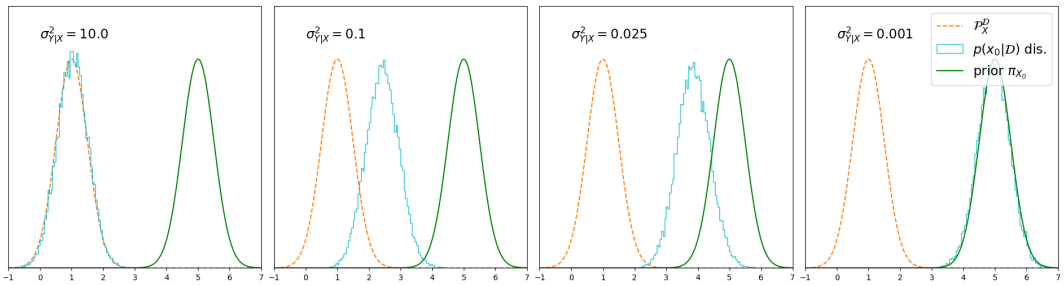


Figure 3.5: Varying degrees of aleatoric uncertainty in the DGP yield the distribution $p(x_0|\mathcal{D})$ to shift between Π_{X_0} and $\mathcal{P}_X^{\mathcal{D}}$.

In this figure, on the one hand, we see that for lower values of $\sigma_{X|Y}$ (the noise standard deviation in the DGP (3.4)) the distribution $p(x_0|\mathcal{D})$ gets closer to Π_{X_0}; while,

on the other hand, this distribution gets closer to $\mathcal{P}_X^{\mathcal{D}}$ for larger values of $\sigma^2_{X|Y}$. This example seems to hint that when the DGP is stained with high (resp. low) aleatoric uncertainty, the distribution $p(x_0|\mathcal{D})$ leans more towards $\mathcal{P}_X^{\mathcal{D}}$ (resp. Π_{X_0}). In section 3.3.4, we provide another example of classification and observe a similar effect.

As we have mentionned before, it therefore follows that the ppd in the discriminative case provides an approximation of (3.28), which leads to a visible mismatch between the ppd and the true (unknown) posterior when the prior Π_{X_0} and $\mathcal{P}_X^{\mathcal{D}}$ are different probability distributions, which we now illustrate. To that end, we compare this PDF to an histogram of samples from the ppd $p(x_0|y_0, \mathcal{D})$ with large $\mathcal{D}$ to remove the effect of epistemic uncertainty and we observe that they perfectly match. Conversely, when Π_{X_0} and $\mathcal{P}_X^{\mathcal{D}}$ are the same distribution, then the ppd indeed approximates the true ppd. This is illustrated in figure 3.6.

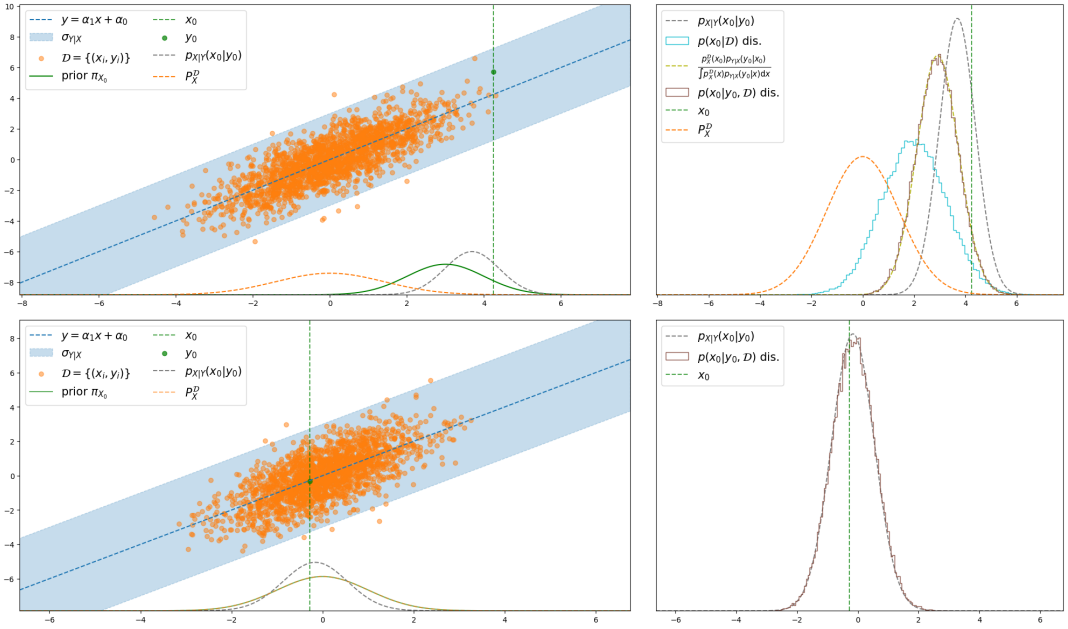


Figure 3.6: The marginal over x_0 is inferred on $\mathcal{D}$: a mismatch between the prior Π_{X_0} and $\mathcal{P}_X^{\mathcal{D}}$ results in a misled approximation (first line). A discriminative approach is accurate in the case where these two distributions match (second line)

We now illustrate a more problematic issue related to the same mechanism. Because the dataset shifts the distribution $p(x_0|\mathcal{D})$, which acts as a prior in the ppd, towards $\mathcal{P}_X^{\mathcal{D}}$ via an implicit density estimation mechanism of $\mathcal{P}_X^{\mathcal{D}}$ with $p(x_0|\mathcal{D})$ the ppd will only assign high probability to the regions of space to which are assigned high probability under $p(x_0|\mathcal{D})$. As a consequence, in the case of affine modeling is that we are not able to predict accurately outside of the support induced by $\mathcal{D}$ as the ppd attributes little to no mass to the true value of x_0. A discriminative affine model cannot extrapolate to regions outside of the support of $\mathcal{D}$ and this conclusion argues, for once, in disfavor of a discriminative approach since an affine model, amongst all models, is expected to extrapolate well. Conversely, as a result of the explicit prior, the generative approach does not suffer from the same shortcoming and we observe that the affine generative

model indeed produces a ppd which assigns high probability to the true value of x_0 and approximates the true unknown posterior. This is illustrated in figure 3.7.

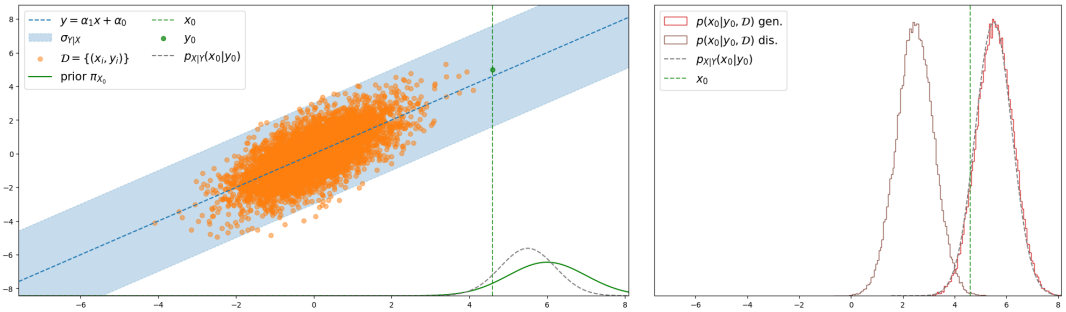


Figure 3.7: Illustration of inaccurate discriminative prediction when outside of the support of $\mathcal{D}$

Classification example

In the previous example of regression, we illustrated how a discriminative modeling approach builds an implicit prior via a density estimation mechanism, resulting in a poor approximation in the case of a mismatch between $\mathcal{P}_X^{\mathcal{D}}$ and the desired prior Π_{X_0}. We also illustrated how, in a discriminative model, the dataset $\mathcal{D}$, the prior Π_{X_0} and the DGP interact to yield $p(x_0|\mathcal{D})$ which, via an implicit density estimation mechanism, is a trade-off between Π_{X_0} and $\mathcal{P}_X^{\mathcal{D}}$. We now also illustrate this specific effect on a classification problem where X is a Categorical variable taking value $c = 1, ..., C$.

We consider a 2-dimensional example with $C = 3$ where the DGP reads: $p_{Y|X=c}(y) = \mathcal{N}(y; r_{Y|X}\left[\mathrm{Re}(\omega^c), \mathrm{Im}(\omega^c)\right]^T, I_2)$ with $\omega = e^{2i\pi/3}$. In this DGP, the aleatoric uncertainty can be controlled via the value of $r_{Y|X}$ which describes the distance of each Gaussian class to the origin ($r\_$ stands for radius). Indeed, the further the different classes are from each other, the easier it is to accurately classify a sample $y \sim \mathcal{P}_{Y_0}$ to its according unknown label. The goal of this section is to illustrate the effect of the roles of distributions $\mathcal{P}_X^{\mathcal{D}}$ and Π_{X_0} in the discriminative approach. We therefore select the two distributions $\mathcal{P}_X^{\mathcal{D}}$ to be different from one another: $\Pr(X = c) = (4 - c)/6$ for $X \sim \Pi_{X_0}$ versus $\Pr(X = c) = c/6$ for $X \sim \mathcal{P}_X^{\mathcal{D}}$ for labels $c = 1, 2, 3$. The distribution $\mathcal{P}_X^{\mathcal{D}}$ defines the frequencies of classes, and together with the DGP can produce a toy dataset $\mathcal{D}$ which is illustrated in the next figure 3.9.

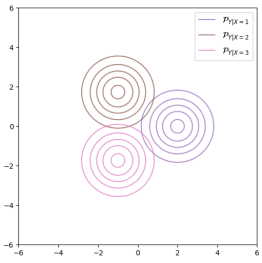


Figure 3.8: Contour plot of the DGP used in the classification example

As a discriminative model, we use a multinomial Logistic classifier with class probability that read:

$$\Pr_\theta(X = c|Y) = \frac{\exp(\beta_c^T Y + \beta_{c,0})}{\sum_{c=1}^{C} \exp(\beta_c^T Y + \beta_{c,0})}, \tag{3.30}$$

where $\theta = \left\{\beta_1, \beta_{1,0}, ..., \beta_C, \beta_{C,0}\right\}$ with $\beta_c, \beta_{c,0} \in \mathbb{R}^2 \times \mathbb{R}$. Similarly to the regression running example, we will display an empirical estimate of distribution $p(x_0|\mathcal{D})$ (3.26)

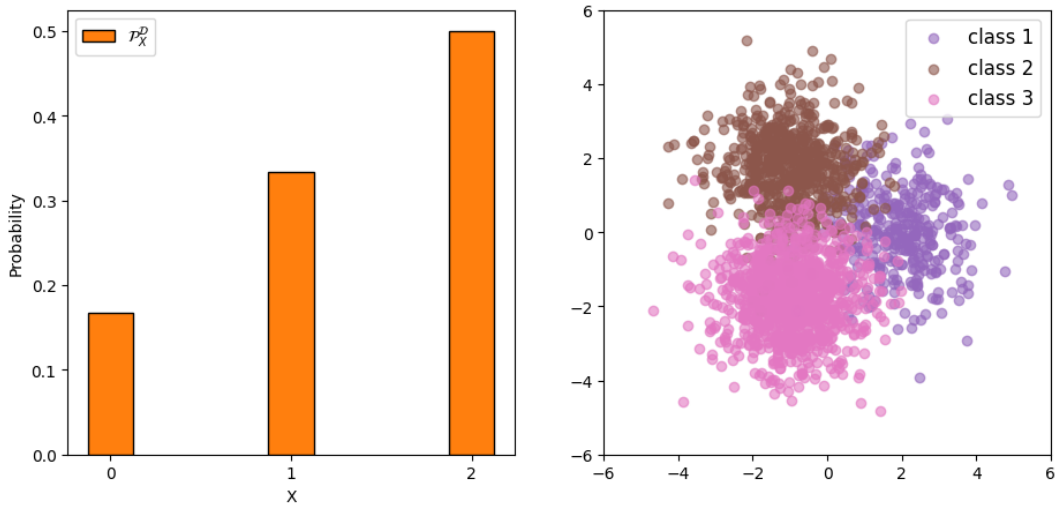


Figure 3.9: Dataset $\mathcal{D}$ with class distribution $\mathcal{P}_X^{\mathcal{D}}$

obtained by the sampling procedure: $y \sim \mathcal{P}_{Y_0}$ (via the unknown DGP which we use only for illustrative purposes), $\theta \sim p(\theta|\mathcal{D})$ and $x_0 \sim \text{Categorical}(\Pr_\theta(X = 1|Y), \Pr_\theta(X = 2|Y), \Pr_\theta(X = 3|Y))$ given by equation (3.30).

A prior distribution over parameters Π_θ which would be conjugate to this logistic model would lead to a posterior $p(\theta|\mathcal{D})$ available in closed form, but unfortunately, such a conjugate prior does not exists amongst the usual probability distributions. We therefore use a simple Gaussian prior over parameter θ and resort to sampling from the ppd using a Metropolis-Hastings MCMC algorithm, though a Gibbs sampling scheme is also available in this setting (41). In the next figure, we display empirical estimates of $p(x_0|\mathcal{D})$ against Π_{X_0} and $\mathcal{P}_X^{\mathcal{D}}$. This figure once again illustrates that the distribution

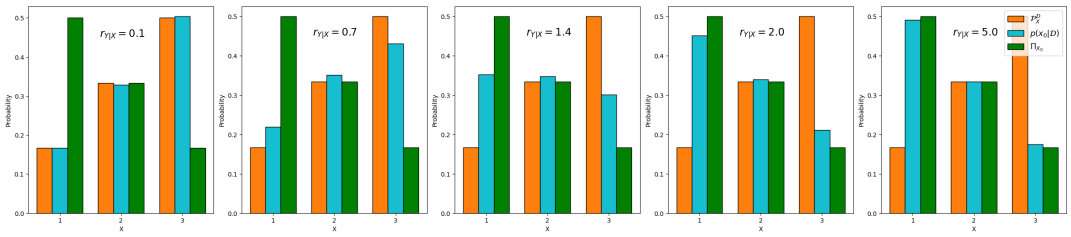


Figure 3.10: Varying degrees of aleatoric uncertainty in the DGP yield the distribution $p(x_0|\mathcal{D})$ to shift between Π_{X_0} and $\mathcal{P}_X^{\mathcal{D}}$.

$p(x_0|\mathcal{D})$ (which we recall acts as a prior in the ppd (3.24)) corresponds to a trade-off between Π_{X_0} and $\mathcal{P}_X^{\mathcal{D}}$. With a very similar interpretation to that of figure 3.5, this figure also seems to hint that the dynamic of the DGP dictates the trade-off between the two distributions: high (resp. low) aleatoric uncertainty shifts $p(x_0|\mathcal{D})$ more towards $\mathcal{P}_X^{\mathcal{D}}$ (resp. Π_{X_0}).

3.3.5 Gibbs sampling from the ppd

We now come to sampling from the ppd. Samples from the ppd can be obtained in two distinct ways. If exact computation of expectation w.r.t. $p(\theta|\mathcal{D})$ is feasible, then (3.22) and (3.23) give tractable expressions (possibly up to a constant) for the ppd. However, the integrals in these equations admit closed form expressions only in specific cases when using conjugate models, and exact integration is, more often than not, unfeasible making the posterior PDF intractable. So in practice, we rather resort to sampling via the joint distribution since, as we explained before, sampling from the joint distribution (3.3) produce samples x_0 that are distributed according to the ppd (3.2).

With that regard, in both modeling cases, the joint PDF, (3.18) and (3.20), can be computed (at least up to a constant), and so we can sample from the joint distribution via the PDF. More conveniently, the discriminative approach yields a specific factorization of its joint PDF (3.20) which enables a sequential sampling procedure with $\theta \sim p(\theta|\mathcal{D})$ and $x_0 \sim p_\theta(x_0|y_0)$. It is therefore motivated to construct models for which the posterior probability distribution of parameters is available in closed form, which can be achieved using conjugacy. By contrast, the generative approach does not induce the same factorization and does not benefit from the same convenient sequential sampling scheme. In this section, we propose a general scheme for sampling from the ppd which can be applied to both modeling approach.

We now propose a scheme that enables to sample from the joint distribution (3.3), and therefore from its x_0 marginal which is nothing but the ppd (3.2). This sampling scheme is based on the notorious Gibbs sampling (11)(30) which is an MCMC algorithm (79) that applies specifically to a joint distribution $p(u, v)$ with a Markov transition $q_t(u^{(t+1)}, v^{(t+1)}|u^{(t)}, v^{(t)}) = p(u^{(t+1)}|v^{(t+1)})p(v^{(t+1)}|u^{(t)})$. This transition leaves the joint distribution $p(u, v)$ invariant since the Gibbs algorithm can be seen as a succession of two steps of Metropolis Hastings (13) transition where the acceptance probability is 1. We apply the principle of Gibbs sampling to the joint distribution $p(x_0, \theta|y_0, \mathcal{D})$ in the generative (resp. discriminative) setting. First, conditionally on the current model $\theta^{(t-1)}$, x_0 is distributed according to the posterior for that model. So the first step of the Markov transition is to draw $x_0^{(t)}$ from equation (3.6) (resp. (3.7)) with $\theta^{(t-1)}$. Then, conditionally on the current value of $x_0^{(t)}$, θ is distributed according to $p(\theta|x_0^{(t)}, y_0, \mathcal{D})$ and so the second step of the markov transition consists in sampling $\theta^{(t)} \sim p(\theta|\mathcal{D}_+^{(t)})$ with the analogous of (3.19) (resp. (3.21)) where $\mathcal{D}_+^{(t)} = \mathcal{D} \cup \{(x_0^{(t)}, y_0)\}$. We summarize this Gibbs sampler in algorithm 3 (for the moment readers should disregard the red parts of the algorithm, as they are related to the semi-supervised learning task covered in section 3.4).

Though this algorithm is written in a similar fashion in both modeling approaches, the conclusions with regard to the different behaviours of the two modeling approaches presented in the previous sections still hold as they are the result of a structural difference between the generative and discriminative approach. This algorithm will be especially useful in the semi-supervised setting, which we describe in the next section 3.4.

In the case of multiple observations $y_{0,1}, ..., y_{0,N_0}$ as in section 3.2.2, the previous algorithm can be effortlessly adjusted in the generative case (recall that the discrimina-

tive case is not compatible with multiple observations, see section 3.2.2). Indeed in this setting, at time t, we first sample $x_0^{(t)}$ according to (3.13) for current model $\theta^{(t-1)}$; and then the dataset $\mathcal{D}$ is augmented into $\mathcal{D}_+^{(t)} = \mathcal{D} \bigcup_{i=1}^{N_0} \{(x_0^{(t)}, y_{0,i})\}$.

Illustrative running example

We now come back to the continued example of affine modeling to provide an example of the Gibbs algorithm mechanism. We assume prior knowledge over parameter θ in the form of $\pi_\Theta(\theta) = \mathcal{N}(\beta; \mu_\beta, \Sigma_\beta)\text{I}\Gamma(\sigma^2; \lambda, \eta)$ where $\mu_\beta \in \mathbb{R}^2$ and $\Sigma_\beta \in \mathbb{R}^{2\times 2}$ is a covariance matrix, IΓ is the PDF associated with an inverse gamma distribution and $\lambda, \eta > 0$ are respectively the shape and scale parameters of the corresponding gamma distribution. Unfortunately, the posterior $p(\theta|\mathcal{D})$ does not admit a closed form expression; but, at least, this choice of conjugate priors allows for both conditionals to be tractable. We first explicit these two conditionals PDF in the generative (resp. discriminative) setting:

$$p(\beta|\sigma^2, \mathcal{D}) \overset{G}{=} \mathcal{N}(\beta; (\frac{\mathbf{X}^T\mathbf{X}}{\sigma^2} + \Sigma_\beta^{-1})^{-1}(\frac{\mathbf{X}^T\mathbf{Y}}{\sigma^2} + \Sigma_\beta\mu_\beta), (\frac{\mathbf{X}^T\mathbf{X}}{\sigma^2} + \Sigma_\beta^{-1})^{-1}), \tag{3.31}$$

$$\text{where } \mathbf{X} \overset{G}{=} \begin{bmatrix} x_1, 1 \\ ... \\ x_{|\mathcal{D}|}, 1 \end{bmatrix}, \mathbf{Y} \overset{G}{=} \begin{bmatrix} y_1 \\ ... \\ y_{|\mathcal{D}|} \end{bmatrix};$$

$$p(\sigma^2|\beta, \mathcal{D}) \overset{G}{=} \text{I}\Gamma(\sigma^2; \lambda + \frac{|\mathcal{D}|}{2}, \eta + \sum_{i=1}^{|\mathcal{D}|} \frac{(y_i - \beta_1 x_i - \beta_0)^2}{2}). \tag{3.32}$$

$$p(\beta|\sigma^2, \mathcal{D}) \overset{D}{=} \mathcal{N}(\beta; (\frac{\mathbf{Y}^T\mathbf{Y}}{\sigma^2} + \Sigma_\beta^{-1})^{-1}(\frac{\mathbf{Y}^T\mathbf{X}}{\sigma^2} + \Sigma_\beta\mu_\beta), (\frac{\mathbf{Y}^T\mathbf{Y}}{\sigma^2} + \Sigma_\beta^{-1})^{-1}), \tag{3.33}$$

$$\text{where } \mathbf{Y} \overset{D}{=} \begin{bmatrix} y_1, 1 \\ ... \\ y_{|\mathcal{D}|}, 1 \end{bmatrix}, \mathbf{X} \overset{D}{=} \begin{bmatrix} x_1 \\ ... \\ x_{|\mathcal{D}|} \end{bmatrix};$$

$$p(\sigma^2|\beta, \mathcal{D}) \overset{D}{=} \text{I}\Gamma(\sigma^2; \lambda + \frac{|\mathcal{D}|}{2}, \eta + \sum_{i=1}^{|\mathcal{D}|} \frac{(x_i - \beta_1 y_i - \beta_0)^2}{2}). \tag{3.34}$$

So in both modeling cases, the posterior distribution $p(\theta|\mathcal{D})$ can be sampled from using a Gibbs scheme by sequentially sampling these two conditionnals. Then, from a Gibbs sampling of affine models, we can almost effortlessly obtain samples from the ppd by including the additional step of sampling x_0 from $p(x_0|y_0, \theta)$ for the current model parameters within the Gibbs sequential Markovian transition. Again, in supplementary material, we summarize this Gibbs sampler in the specific case of affine homoskedastic modelling, see algorithm 4 (readers should disegard the steps in red for now, as they are related to semi-supervised learning which we now discuss).

3.4 Bayesian Semi-Supervised learning

In this section, we now build upon the equations, arguments and conclusions presented in the previous sections to tackle the problem of semi-supervised learning. As we have mentionned before, supervised learning techniques use the observed variables and their corresponding labels to build a model which capture the dependency between two rv and which can be used to make predictions about the label conditionally on the value of observed rv. Conversely, unsupervised learning (40)(38) take interest in learning pattern in a data distribution without considering the notion of associated labels. Structure can be represented by data clusters obtained by K-means (55)(54), graph-based clustering methods such as spectral clustering (67) or Louvain method (7), or via maximum-likelihood in a mixture probability distribution model (18). In the beginning of this chapter, we explained that the Generative approach for modeling the unknown posterior, should not be confused with the tasks and techniques of Generative modeling. These techniques can also be considered as unsupervised learning as they enable to capture the structure from univariate data such that the corresponding probabilistic model can be sampled from easily in order to obtain observations which are approximately distributed according to the dataset distribution. Most popular methods include Variational AutoEncoders (49), Generative Adversarial Networks (33), Normalizing Flows (70) and Diffusion models (81) but this is beyond that we consider in this chapter. Semi-supervised learning does however lie within our scope as it aims to obtain a conditional model to predict label from observations, but the goal is to infer the model from both a labeled dataset and unlabeled observations. In this section we now build upon the previous arguments and discuss the compatibility of both learning approach with this learning task.

3.4.1 The learning task

In section 3.2 and onward, we presented the general task of learning a model for the posterior (3.1) using a set of *labeled* observations $\mathcal{D}$, and how to predict about an x_0 given a corresponding observation y_0 with the ppd, which we now refer to as a supervised learning task. However, in many statistical learning settings, we also dispose of *unlabeled* observations. They corresponds to values $\widetilde{y}_j$, which we know (or assume) are produced by the DGP, but for an unknown values $\widetilde{x}_j$ for which we assume prior knowledge $\Pi_{\widetilde{X}_j}$: $\mathcal{Y} = \{\widetilde{y}_j | \exists \widetilde{x}_j \sim \Pi_{\widetilde{X}_j}, \widetilde{y}_j \sim \mathcal{P}_{Y|X}(Y|X = \widetilde{x}_j)\}_{j=1}^{|\mathcal{Y}|}$. When the observations in $\mathcal{D}$ and $\mathcal{Y}$ cover different regions of the observation space, and/or when $\mathcal{Y}$ has a significant amount of elements, then the unlabeled observations $\mathcal{Y}$ may contain significant or non-negligible information (68). In this context, a semi-supervised learning task aims at inferring a model from both labeled and unlabeled observations. This question has risen in importance in importance where we dispose of a lot of unlabeled observations, but where the labelling tasks is expensive (as it is the case when the labelling needs to be conducted by a human operator).

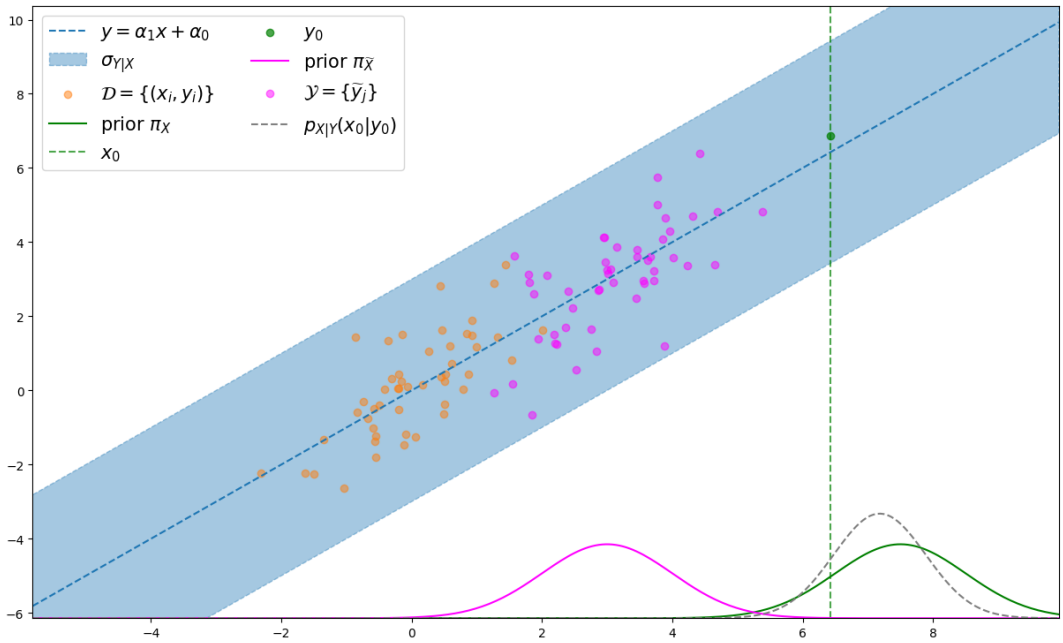


Figure 3.11: Affine semi-supervised regression setting

The ppd (3.2) then becomes:

$$p(x_0|y_0, \mathcal{D}, \mathcal{Y}) = \int_\Theta p(x_0|y_0, \theta) p(\theta|y_0, \mathcal{D}, \mathcal{Y}) \mathrm{d}\theta, \tag{3.35}$$

and we aim to compute this PDF, or sample this distribution if exact computation of the integral is not feasible. This formulation is more general than the one described in section 3.2 and it reduces to supervised learning in the case where $\mathcal{Y} = \varnothing$.

Throughout this section, we will carry on using the continued example of affine modeling to illustrate the arguments and conclusions. To that end, we suppose that, in addition to $\mathcal{D}$, we also dispose of unlabeled observations $\widetilde{y}_j$ produced from the DGP (3.4) via an unknown label $\widetilde{x}_j$ for which we suppose prior knowledge in the form of a prior $\Pi_{\widetilde{X}}$ which is supposed to be the same for all $j = 1, ..., |\mathcal{Y}|$ and which we consider to be Gaussian $\pi_{\widetilde{X}}(\widetilde{x}_j) = \mathcal{N}(\widetilde{x}_j; \mu_{\widetilde{X}}, \sigma^2_{\widetilde{X}})$. The semi-supervised setting is illustrated in figure 3.11.

3.4.2 Both modeling confronted to the semi-supervised learning task

We now confront the two modeling approaches to the specific problem of semi-supervised learning by analysing the model posterior which reads:

$$p(\theta|y_0, \mathcal{D}, \mathcal{Y}) = p(\theta|\mathcal{D}) \frac{p(y_0|\theta) p(\mathcal{Y}|\theta)}{p(y_0, \mathcal{Y}|\mathcal{D})}. \tag{3.36}$$

We first explain that the discriminative approach does not allow for Bayesian semi-supervised learning. To see this, recall the conclusion of section 3.3.1: when we do not

know x_0, the posterior over θ does not depend on y_0 and so this observation does not carry any information to the discriminative models. Therefore, the same applies for the elements of $\mathcal{Y}$: since we do not know the label $\widetilde{x}_j$, the unlabeled observation does not bring any information on θ as the posterior over models does not depend on $\widetilde{y}_j$. Finally, the model posterior (3.36) reduces to $p(\theta|\mathcal{D})$, (3.35) reduces to (3.23) and all the other equations in section concerning the discriminative modeling approach remain unchanged.

Conversely, the generative approach indeed allows for semi-supervised learning. In section 3.3.1, we explained that, even though we do not know the value of x_0, the posterior distribution over models still depends on the observation y_0 indirectly through the prior Π_{X_0}. With a similar argument, we understand that the unlabeled data $\mathcal{Y}$ indeed carry information on model θ. We write: $p(\mathcal{Y}|\theta) \overset{G}{=} \prod_{j=1}^{|\mathcal{Y}|} \int p_\theta(\widetilde{y}_j|\widetilde{x}_j)\pi_{\widetilde{X}_j}(\widetilde{x}_j)\mathrm{d}\widetilde{x}_j$. So, even though we do not know the label $\widetilde{x}_j$, probable generative models θ under (3.36) produce, with high probability, the value of $\widetilde{y}_j$ for some unknown label distributed under the prior $\Pi_{\widetilde{X}_j}$. In a classification setting, this term can indeed be computed as a tractable finite sum (4)(45), but in general, this term is only available in integral form making the joint PDF $p(x_0, \theta|y_0, \mathcal{D}, \mathcal{Y}) \overset{G}{\propto} \pi_{X_0}(x_0)p_\theta(y_0|x_0)p(\theta|\mathcal{D})p(\mathcal{Y}|\theta)$, intractable, not even up to a constant. This raises the question of sampling from the ppd since its PDF is intractable. In the next section, we propose to use a variation of the Gibbs algorithm presented in section 3.3.5, and which allows us to sample from this ppd.

3.4.3 A Gibbs sampling algorithm for semi-supervised learning

In this section we extend the previous Gibbs sampling algorithm presented in section 3.3.5 for sampling from the ppd (3.35) in the case of generative semi-supervised learning. To that end, we apply the Gibbs mechanism of sequentially sampling the conditional distributions in the joint distribution $p(x_0, \widetilde{x}_1, ..., \widetilde{x}_{|\mathcal{Y}|}, \theta|y_0, \mathcal{D}, \mathcal{Y})$. Firstly, conditionally on the current value of $\theta^{(t-1)}$, the labels $x_0, \widetilde{x}_1, ..., \widetilde{x}_{|\mathcal{Y}|}$ are independent and each distributed according to its own posterior distribution. So the first step of the Gibbs Markovian transition is to sample $x_0^{(t)} \sim p(x_0|y_0, \theta^{(t-1)})$ with equation (3.6) (as in section 3.3.5) and $\widetilde{x}_j^{(t)} \sim p(\widetilde{x}_j|\widetilde{y}_j, \theta^{(t-1)})$. Secondly, conditionally on the current label values $x_0^{(t)}, \widetilde{x}_1^{(t)}, ..., \widetilde{x}_{|\mathcal{Y}|}^{(t)}$, the model parameters are distributed according to $p(\theta|\mathcal{D}, x_0^{(t)}, y_0, \widetilde{x}_1^{(t)}, \widetilde{y}_1, ..., \widetilde{x}_{|\mathcal{Y}|}^{(t)}, \widetilde{y}_{|\mathcal{Y}|})$. So the second step of the Gibbs Markovian transition is to sample $\theta^{(t)} \sim p(\theta|\mathcal{D}_+^{(t)})$ analogous of equation (3.19) where $\mathcal{D}_+^{(t)} = \mathcal{D} \cup (x_0^{(t)}, y_0) \cup \{(\widetilde{x}_j^{(t)}, \widetilde{y}_j)\}_{j=1}^{|\mathcal{Y}|}$. We summarize this Gibbs mechanism in algorithm 3 and we highlight in red the steps which are effectively responsible for semi-supervised learning.

In the case of semi-supervised learning setting, this Gibbs algorithm is all the more crucial. Indeed, while in the supervised context the Gibbs approach was only an alternative option to sampling from the joint distribution with a PDF (3.18) which could be computed up to a constant; in the case of semi-supervised learning however, it is possible that this joint PDF cannot be evaluated, not even up to a constant and the Gibbs approach is therefore very convenient for sampling the corresponding ppd.

We have written the Gibbs algorithm with $\mathcal{Y}$ included in the inference to perform

Algorithm 3 Gibbs sampling from $p(x_0|y_0, \mathcal{D}, \mathcal{Y})$ in the generative (resp. discriminative) setting

Require: $y_0, \mathcal{D}, \mathcal{Y}$, number of steps T
  $\theta^{(0)} \sim p(\theta|\mathcal{D})$
  for $t = 1$ to T **do**
    $x_0^{(t)} \sim p(x_0|y_0, \theta^{(t-1)})$ with (3.6) (resp. (3.7)
    for all $\widetilde{y}_j \in \mathcal{Y}$ **do**
      sample $\widetilde{x}_j^{(t)} \sim p(\widetilde{x}_j|\widetilde{y}_j, \theta^{(t-1)}) = \frac{p_{\theta^{(t-1)}}(\widetilde{y}_j|\widetilde{x}_j)\pi_{\widetilde{X}_j}(\widetilde{x}_j)}{p(\widetilde{y}_j|\theta^{(t-1)})}$
    end for
    set $\mathcal{D}_+^{(t)} = \mathcal{D} \cup (x_0^{(t)}, y_0) \cup \{(\widetilde{x}_j^{(t)}, \widetilde{y}_j)\}_{j=1}^{|\mathcal{Y}|}$ and $\theta^{(t)} \sim p(\theta|\mathcal{D}_+^{(t)})$ with (3.19) (resp. (3.21))
  end for
  return $x_0^{(T)}$

semi-supervised learning for both modeling approach but of course, as we have mentionned before, the semi-supervised ppd (3.35) reduces to the supervised ppd (3.2) in the discriminative setting, so this Gibbs algorithm, even though it involves $\mathcal{Y}$, is not able to leverage any information from the unlabeled observations in this modeling approach. We therefore would like to stress that the semi-supervised learning is not enabled by the Gibbs procedure itself but rather by using a generative modeling instead of a discriminative one which induces different conditional dependency between all the rv. We proposed the Gibbs algorithm as a way to sample the joint distributions $p(x_0, \theta|y_0, \mathcal{D}, \mathcal{Y})$ in the case of generative modeling (with possibly $\mathcal{D} = \varnothing$ in the supervised setting) which is particularly convenient to use because of the conditional independence (w.r.t. θ) of labels $x_0, \widetilde{x}_1, ..., \widetilde{x}_{|\mathcal{Y}|}$. However, this Gibbs scheme it is only a possible approach for sampling from the corresponding ppd which effectively depends on $\mathcal{Y}$.

We now come back to the continued example of affine modeling and use it to illustrate the practical use of the Gibbs sampling algorithm which we use to illustrate, respectively, the compatibility and incompatibility between the generative and discriminative approaches and the semi-supervised learning. In algorithm SM3.1 we first explicit how to build upon the supervised Gibbs sampling algorithm applied to the supervised ppd presented in section 3.3.5 to obtain a Gibbs scheme in order to sample from the semi-supervised ppd.

In this case, the generative and discriminative approach yield two different equations for the posterior $p(\widetilde{x}_j|\widetilde{y}_j, \theta)$:

$$p(\widetilde{x}_j|\widetilde{y}_j, \beta, \sigma^2) \overset{G}{=} \mathcal{N}(\widetilde{x}_j; (\frac{1}{\sigma^2} + \frac{1}{\sigma_{\widetilde{X}}^2})^{-1}(\frac{\beta_1(\widetilde{y}_j - \beta_0)}{\sigma^2} + \frac{\mu_{\widetilde{X}}}{\sigma_{\widetilde{X}}^2}), (\frac{1}{\sigma^2} + \frac{1}{\sigma_{\widetilde{X}}^2})^{-1}), \tag{3.37}$$

$$p(\widetilde{x}_j|\widetilde{y}_j, \beta, \sigma^2) \overset{D}{=} \mathcal{N}(\widetilde{x}_j; \beta_1\widetilde{y}_j + \beta_0, \sigma^2). \tag{3.38}$$

The Gibbs sampling procedure for semi-supervised learning of affine homoskedastic is summarized in algorithm 4. Again, the steps highlighted in red are indeed responsible for leveraging information from the unlabeled observations.

We will now illustrate whether or not the modeling approach enables leveraging information in $\mathcal{Y}$ to contribute in reducing the epistemic uncertainty. We first provide

Algorithm 4 Gibbs sampling from $p(x_0|y_0, \mathcal{D}, \mathcal{Y})$ using a generative (resp. discriminative) homoskedastic affine model

Require: $y_0, \mathcal{D}$, $\mathcal{Y}$, number of steps T
  $\sigma^{2(0)} \sim \text{I}\Gamma(\sigma^2; \lambda, \eta)$
  $\beta^{(0)} \sim p(\beta|\sigma^{2(0)}, \mathcal{D})$ with (3.31)
  for $t = 1$ to T **do**
    $x_0^{(t)} \sim p(x_0|y_0, \beta^{(t-1)}, \sigma^{2(t-1)})$ with (3.10) (resp. (3.11)
    for all $\widetilde{y}_j \in \mathcal{Y}$ **do**
      $\widetilde{x}_j^{(t)} \sim p(\widetilde{x}_j|\widetilde{y}_j, \beta^{(t-1)}, \sigma^{2(t-1)})$ with (3.37) (resp. (3.38))
    end for
    set $\mathcal{D}_+^{(t)} = \mathcal{D} \cup (x_0^{(t)}, y_0) \cup \{(\widetilde{x}_j^{(t)}, \widetilde{y}_j)\}_{j=1}^{|\mathcal{Y}|}$, $\sigma^{2(t)} \sim p(\sigma^2|\beta^{(t-1)}, \mathcal{D}_+^{(t)})$ with (3.32) (resp. (3.34))
    $\beta^{(t)} \sim p(\beta|\sigma^{2(t)}, \mathcal{D}_+^{(t)})$ with (3.31)(resp. (3.33))
  end for
  return $x_0^{(T)}$

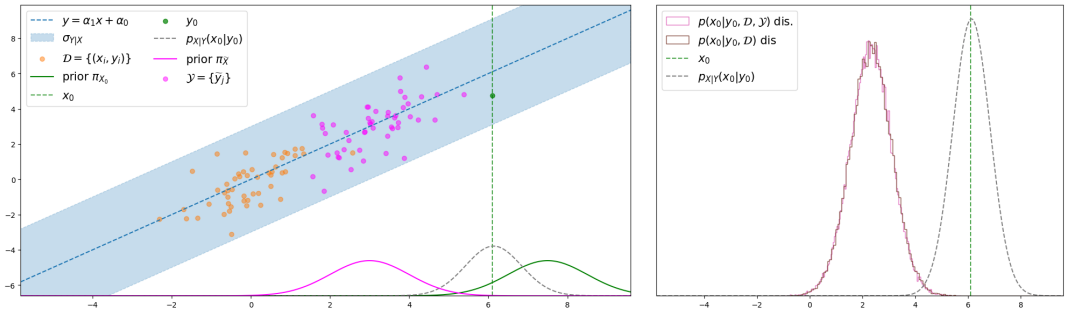
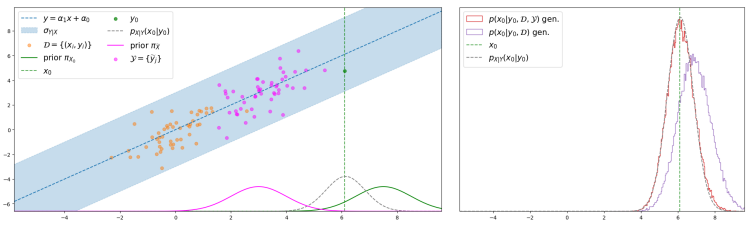


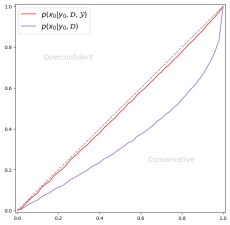
Figure 3.12: Samples from the supervised and semi-supervised Gibbs algorithm in affine discriminative modeling

empirical evidence that the discriminative approach is unable to leverage information in unlabeled observations to reduce the epistemic uncertainty. To that end, we consider a setting where $\mathcal{D}$ contains few points leading to high epistemic uncertainty to be reduced; and we apply the previous Gibbs sampling algorithm SM3.1 in the discriminative case with and without unlabeled observations $\mathcal{Y}$ and visually observe that the resulting samples seem to follow the same distribution. This empirical illustration is presented in figure 3.12. We further perform a Kolmogorov-Smirnov statistical test where the null hypothesis is H_0 : the samples obtained from the supervised and semi-supervised Gibbs sampling algorithm are from the same (unknown) distribution. This test works from two sets of iid samples; but the samples from the Gibbs sampling algorithm are correlated samples so we first extract (almost) uncorrelated samples sub-sampling the Markov chain at every each integrated auto-correlation time steps. The Kolmogorov-Smirnov test yields a p-value of 0.566 and we cannot reject the null hypothesis with low error probability; which indicates that the data i.e. the two sets of de-correlated samples is consistent with the null hypothesis i.e. that they originate from the same underlying probability distribution.

Conversely, in the generative modeling approach, we go back to the setting presented in the previous figure 3.11, and we compare the supervised and semi-supervised ppds, from which we obtain samples via the corresponding Gibbs sampling algorithm 4 with and without $\mathcal{Y}$. The results are presented in the next figure 3.13a. We compare the empirical distributions (built an histogram of the samples) and we visually observe that the semi-supervised ppd provides with a better of the true unknown posterior than the supervised ppd. To go further than a visual interpretation, we compare the calibration curves of the both ppds. The calibration curve allows to assess the quality of an approximation and is computed from the PDF of the target posterior, in our case the true unknown posterior and from samples from the approximating distribution, in our case the (supervised or semi-supervised) ppds. It is built by computing, for values $\alpha \in [0, 1]$, the α-highest density region of the target distribution using the PDF and computing the proportion of samples which land in that region. We can therefore conclude that the unlabeled observation were taken into account during the inference and indeed contributed to reducing the epistemic uncertainty.



(a) Unlabeled observations $\mathcal{Y}$ indeed contribute to reducing the generative epistemic uncertainty.



(b) Calibration curves of supervised and semi-supervised ppds.

Figure 3.13: Semi-supervised generative affine modeling

3.4.4 Parallel inference

In this section we propose to re-discuss the problems of supervised and semi-supervised learning by considering them as solving two (or several) posterior inferences at the same time. To that end, we now denote $\mathcal{X} = \{\widetilde{x}_j | \widetilde{y}_j \sim \mathcal{P}_{Y|X}(Y|X = \widetilde{x}_j\}_{j=1}^{|\mathcal{Y}|}$ the set of unknown labels associated with the unlabeled observations. On the one hand, $\mathcal{X}$ are not necessarily label values of interest in the initial problem (that of inferring x_0 via y_0), it is nonetheless an unknown rv related to $\mathcal{Y}$ via the same unknown DGP and can be associated with an inference problem (again, possibly irrelevant in the context of the initial problem). On the other hand, both x_0 and $\mathcal{X}$ can be values of interest. Indeed, in many learning instances, we dispose of a training dataset (the set $\mathcal{D}$) to infer the probable models; and given another set of observations (the so-called testing dataset), our aim is to predict for each of them the associated label (which is different from one observation to another, as opposed to section 3.2.2, where one common label produced several observations). In this case, all the observations in the testing dataset play an epistemic role in the generative case. More precisely, the observations of the testing dataset act

as unlabeled observations resulting in an underlying semi-supervised learning setting. This is illustrated via quantitative simulations in the next section 3.5.

In this context, since we are trying to infer x_0 from y_0 and $\mathcal{X}$ from $\mathcal{Y}$ via probable models θ using the same observations $\mathcal{D}$, the two inferences should not be treated as independent problems in both modeling approaches. However, a main difference between the two modeling approaches can be understood when considering two (or several) inference problems at once.

In the discriminative setting, we have $p(x_0, \mathcal{X}|y_0, \mathcal{Y}, \mathcal{D}) \overset{D}{=} p(x_0|y_0, \mathcal{D})p(\mathcal{X}|\mathcal{Y}, \mathcal{D})$ so the inference of x_0 (resp $\mathcal{X}$) does not depend on $\mathcal{Y}$ (resp y_0). As such, each inference problem can be solved via sampling from its corresponding ppd. Conversely, in the generative setting, all the observations $y_0, \mathcal{Y}$ act as unlabelled observations in both inferences and the two problems should not be treated independently as all the unlabeled observations contribute to reducing the epistemic uncertainty. As a result, $p(x_0, \mathcal{X}|y_0, \mathcal{Y}, \mathcal{D}) \overset{G}{\neq} p(x_0|y_0, \mathcal{D})p(\mathcal{X}|\mathcal{Y}, \mathcal{D})$ and both inference can not be solved by sampling from each corresponding ppd, and this modelling approach instead calls for a sampling from the joint distribution. Finally, note that the previous semi-supervised algorithm, which we proposed as a way to obtain samples from the posterior $p(x_0|y_0, \mathcal{D}, \mathcal{Y})$ was constructed by applying the Gibbs sequential sampling mechanism to the joint distribution $p(x_0, \mathcal{X}, \theta|y_0, \mathcal{Y}, \mathcal{D})$ and as such, produced desired samples from $p(x_0|y_0, \mathcal{D}, \mathcal{Y})$ but also, as a byproduct, samples from $p(\mathcal{X}|\mathcal{Y}, \mathcal{D}, y_0)$, effectively solving both inference problems at once. Again of course, since the Gibbs is only a tool for solving the inference problem, the underlying structure of dependency between rv is preserved. So in the discriminative case these distributions respectively reduce to $p(x_0|y_0, \mathcal{D})$ and $p(\mathcal{X}|\mathcal{Y}, \mathcal{D})$.

3.5 Simulations

Throughout the chapter, we leveraged the example of affine homoskedastic modeling in the case of univariate regression to illustrate the arguments. As we mentionned before, this illustrating example can be relevant to some readers as modeling affine dependencies is a frequent problem in many scientific fields. We also used this example as it enables, with appropriate choice of priors, to use a straightforward Gibbs sampler in both generative and discriminative modeling approaches. However, considering models which enable such convenient sampling procedures with closed form $p(\theta|\mathcal{D})$ (or all its conditionals in a Gibbs scheme) indeed heavily restricts the choice of model $\mathcal{P}_\theta$. In this section we now consider conditional models $\mathcal{P}_\theta$ defined using NN functions with tractable PDF, but for which we are not able to elicit a prior Π_θ over parameters θ such that the posterior distribution $p(\theta|\mathcal{D})$ admits a closed form expression, and we resort to approximate sampling from that posterior using Stochastic Gradient Langevin Dynamics (86).

In this section, we tackle the problem of classification in which X is hence a Categorical rv. We evaluate generative and discriminative models both defined via NN functions (we describe the specific structure hereafter) and with a similar number of parameters. We proceed to assess the classification accuracy of a generative versus discriminative model for $\theta \sim p(\theta|\mathcal{D}, y_0)$ via Gibbs sampling. We consider three different scenarios

which we now describe.

Following the idea described in section 3.4.4, we now consider two distincts sets: the training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, and a testing dataset $\{(x_{0,j}, y_{0,j})\}_{j=1}^M$. All the labels $x_{0,j}$ are distributed according to Π_{X_0}, and in the following we address three scenarios, which differ in the size of the testing set, and the possible discrepancy between Π_{X_0} and $\mathcal{P}_X^{\mathcal{D}}$.

Scenario 1: Identical priors and sizes. We first consider the scenario where the label distribution from the training dataset coincides with the prior. So $\mathcal{P}_X^{\mathcal{D}} = \Pi_{X_0}$, and as such, couples (x_i, y_i) (which belong to training dataset $\mathcal{D}$) and $(x_{0,j}, y_{0,j})$ (which belong to the testing dataset) have the same distributions. This corresponds to the most frequent situation in practice and we use this setting as a baseline. We consider the dataset $\mathcal{D}$ and the set $\{y_{0,j}\}$ of unlabeled observations to be of the same size, i.e. $N = M$.

Scenario 2: Imbalanced dataset (different priors, same sizes). We then consider the scenario where $\mathcal{P}_X^{\mathcal{D}} \neq \Pi_{X_0}$, so couples (x_i, y_i) and $(x_{0,j}, y_{0,j})$ do not have the same distributions (and indeed are quite different - see figure 3.14). We still set $N = M$.

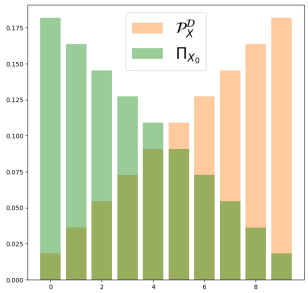


Figure 3.14: Different prior distributions in scenario 2.

Scenario 3: Few labeled samples (same priors, different sizes). We finally consider the scenario where $\mathcal{P}_X^{\mathcal{D}} = \Pi_{X_0}$, but $N \ll M$, so we dispose of a few labeled observations in $\mathcal{D}$, and of a large amount of observed values $y_{0,j}$ for which we want to infer the corresponding label $x_{0,j}$. In this setting, the low number of observations in $\mathcal{D}$ will hinder the prediction accuracy of both models, but as we have explained before, the large amount of unlabeled observations act as an unlabeled dataset in the generative case.

We consider both the classification datasets of MNIST and of FashionMNIST, for which we reduce the dimension of observations via a Principal Component Analysis (73) in order to keep 95% of explained variation. We compare a discriminative model which is a fully connected NN with 4 hidden-layers of 256 units to a generative model which is built using a combination of invertible conditional Normalizing Flows layers (21) and stochastic ones (3). We sample from the joint distribution via Gibbs sampling with $T = 10$ steps. The results are provided in the next table 3.1 and for each dataset and scenario, we consider 10 independent runs and we display the average classification accuracy as well as the standard deviation.

We now analyze the results of this experiment. Comparing the results for the first scenario tells us that the generative modeling can indeed reach similar classification accuracy to its discriminative counterpart. This scenario can be used as a baseline experiment for the two following scenarios. In the second scenario, the distribution of labels in the dataset, $\mathcal{P}_X^{\mathcal{D}}$, is different from Π_{X_0}, leading to a situation of imbalanced dataset. In this situation, we notice that the discriminative model indeed suffers in term of accuracy as it favors the dominant classes of the dataset. As we have explained, the generative approach does not suffer from such dataset imbalance, or at least not

Dataset	MNIST		FashionMNIST	
Model.	Disc.	Gen.	Disc.	Gen.
Scenario 1	0.9628 ± 0.0014	$\mathbf{0.9749 \pm 0.0145}$	$\mathbf{0.8767 \pm 0.0009}$	0.8652 ± 0.0094
Scenario 2	0.9520 ± 0.0007	$\mathbf{0.9774 \pm 0.0213}$	0.7000 ± 0.0057	$\mathbf{0.8482 \pm 0.0172}$
Scenario 3	0.9349 ± 0.0018	$\mathbf{0.9690 \pm 0.0237}$	0.7618 ± 0.0008	$\mathbf{0.8365 \pm 0.0187}$

Table 3.1: Classification accuracy in percentages of Discriminative and Generative modeling approaches

as much, which is confirmed in this experiment. This experiment is coherent with the conclusions of the discussion in section 3.3.4. Finally, in the third scenario, the lower number of labeled observation in $\mathcal{D}$ hinders, as expected, the classification accuracy of the discriminative model as compared to its generative counterpart, as the latter indeed leverages the unlabeled observations in the inference of probable models, which indeed contributes to reduce the modeling epistemic uncertainty, as discussed in section 3.4.2.

3.6 Conclusion

Throughout this chapter, we discussed Bayesian epistemic UQ in posterior learning tasks using generative and discriminative models. We thus analyzed the PPD and drew several conclusions.

On the one hand, discriminative models are an easy-to-use tool since they can be parameterized easily and directly approximate the posterior. Moreover, if they can be sampled from easily, then one can use a straightforward two-step procedure for sampling from the PPD, which indeed enables quantifying the epistemic uncertainty. However, by nature discriminative models do not take into account the information contained in the prior distribution, which is replaced by an implicit prior inferred on the dataset. As a result, they suffer from imbalanced datasets. Finally they cannot be conveniently used in the context of inferring from multiple observations, and they cannot leverage information from unlabeled data.

On the other hand, generative models are perhaps less convenient to use as they usually require a more sophisticated structure and require an additional inference step, in addition to the prior distribution, to sample from the corresponding posterior. Yet, by construction, they do enable to leverage information from all available sources, making them an appealing tool, in particular in a semi-supervised context. In practice, the two-step procedure for sampling from the PPD is no longer available; but our general purpose Gibbs sampling based algorithm indeed enables to sample from this distribution of interest while taking into account prior knowledge, multiple observations and both labeled and unlabeled datasets.

3.7 Perspectives for future work

In this work, we have tackled epistemic uncertainty quantification using the Bayesian principles and compared generative and discriminative methods under the scope of the PPD. In this context, we have assumed that the observations are produced from the labels by a DPG. We considered the case where the DPG has an untractable PDF, making the PDF of the corresponding distribution of labels of interest given the observations also untractable. As we have mentioned, this situation arises in at least two cases: (i) when the DGP is simply unknown, or (ii) when the DGP is available via its sampling procedure but its PDF is implicit. In this work, we have not specifically considered the second setting, as, unless for illustration purposes, we have not resorted to sampling from the DGP.

Generative versus discriminative active learning and PPD-based acquisition ?

However, in many situations, especially when it comes to likelihood-free (simulation-based) inference settings, the DGP can indeed be used via its simulation procedure and remains available for a practitioner to sample observations from given (or perhaps specifically chosen) labels. In this case, it follows that: (i) ABC methods are theoretically feasible; and (ii) it is possible to specifically choose labels in order to augment the dataset and reduce the epistemic uncertainty (72). We now consider the second point.

Many different acquisition rules for selecting the next label with which to augment the dataset have been proposed, see namely (22)(56)(36). However, a criterion which is motivated by, and based on, the PPD in an uncertainty-aware inference has not yet been proposed. In this regard, several aspects might be topics for future work.

A relevant goal would first be that of obtaining a relevant acquisition criterion, which would instead be centered around the PPD, the same objective considered in the Bayesian posterior learning setting. A first idea could be to consider the following expected information gain criterion:

$$x^* \in \arg\max_x \mathbb{E}_{y\sim\mathcal{P}_{Y|X=x}}\left[D_{\mathrm{KL}}\Big(p(x_0|y_0,\mathcal{D})||p(x_0|y_0,\mathcal{D}\cup\big\{x,y\big\}\Big)\right]. \tag{3.39}$$

This criterion can be interpreted as identifying a value x^* that causes the most significant change in the PPD (in the sense of the Kullback Leibler divergence) upon incorporating in the inference the new couple $\big\{x^*, y\big\}$ where y is the average observation produced by x^* via the DGP. In practice, however, this criterion is difficult to compute or approximate. Therefore, it would be particularly interesting to explore the feasibility of obtaining a suitable approximation of this objective function, or at least a suitable approximation of its gradient with respect to x, thus allowing for an approximate solution using gradient ascent.

- Can we elicit a PPD-based acquisition criterion which can indeed be computed and optimized in practice in an active learning setting?

- More generally, can further differences between generative and discriminative modeling methods be witnessed when considering the active learning setting?

Sequential inference in active learning setting

The task of augmenting the dataset during the inference is often referred to as *active learning* in the literature. In this context, we obtain a sequential problem where (i) we wish to estimate labels and/or model parameters using the current dataset, say $\mathcal{D}_t$, and (ii) we then use the current estimation (and some acquisition rule) to obtain a new input for the DGP, which produces a couple with which to augment the dataset into $\mathcal{D}_{t+1}$. In this context, Bayesian posterior learning (step (i)) becomes a sequential problem of sampling from the sequence of PPDs $\left\{p(x_0|y_0,\mathcal{D}_t)\right\}_{t=1,...}$ or equivalently $\left\{p(x_0\theta|y_0,\mathcal{D}_t)\right\}_{t=1,...}$. In this context, it would indeed be interesting to consider using Sequential Monte Carlo (SMC) algorithms. SMC samplers (see (10) and (14) for relevant references) leverage the principles of invariant Markov transition as in MCMC techniques (we denote $\mathcal{M}_t$ such a $\mathcal{P}_t$-invariant Markov kernel); and the principle of importance-weighting and resampling as in Sampling Importance Resampling. The combination of these two actions allows to rejuvenate and re-use the promising particles with high importance weights and discard the others in order to inductively construct empirical approximations of the target distributions (34). We very briefly describe the principle of Sequential Monte Carlo samplers.

Suppose at time $t-1$, I dispose of the weighted particle-based empirical approximation of the probability measure $\mathcal{P}_{t-1}(\mathrm{d}x) \approx \sum_{i=1}^N w_{t-1}^{(i)} \delta_{X_{t-1}^{(i)}}(\mathrm{d}x)$. The underlying principle of SMC sampling is based on the equation:

$$\mathcal{P}_t(\mathrm{d}x) = \frac{p_t(x)}{p_{t-1}(x)}\mathcal{P}_{t-1}(\mathrm{d}x) \propto \frac{\widetilde{p}_t(x)}{\widetilde{p}_{t-1}(x)}\mathcal{P}_{t-1}(\mathrm{d}x); \tag{3.40}$$

which indeed relates two consecutive measures via the ratio (up to a constant) of PDF. This equation, together with considering $\mathcal{M}_t$ a $\mathcal{P}_t$-invariant Markov kernel, enables us to rewrite:

$$\mathcal{P}_t(\mathrm{d}x) \propto \int \frac{\widetilde{p}_t(x')}{\widetilde{p}_{t-1}(x')}\mathcal{P}_{t-1}(\mathrm{d}x')\mathcal{M}_t(x',\mathrm{d}x). \tag{3.41}$$

Finally, one can obtain a weighted particle-based approximation of the measure at time t, $\mathcal{P}_t(\mathrm{d}x) \approx \sum_{i=1}^N w_t^{(i)} \delta_{X_t^{(i)}}(\mathrm{d}x)$ by plugging in this expression the previous empirical approximation of the measure and by sampling the corresponding expression. This reduces to (i) resampling according to the weights, (ii) propagating the particles according to $\mathcal{M}_t$ and (iii) reweighting the particles:

$$X_t^{(i)} \sim \mathcal{M}_t(X_{t-1}^{(A_t^{(i)})},.)\text{(ii)} \quad \text{where } A_t^{(i)} \sim \text{Categorical}(w_{t-1}^{(0)},...,w_{t-1}^{(N)})\text{(i)} \tag{3.42}$$

$$w_t^{(i)} = \frac{\widetilde{p}_t(X_t^{(i)})/\widetilde{p}_{t-1}(X_t^{(i)})}{\sum_{j=1}^N \widetilde{p}_t(X_t^{(j)})/\widetilde{p}_{t-1}(X_t^{(j)})}\text{(iii)}. \tag{3.43}$$

In this procedure, it is not necessary to proceed to resampling at each step, in which case the steps read:

$$X_t^{(i)} \sim \mathcal{M}_t(X_{t-1}^{(i)},.) \text{ and } w_t^{(i)} = \frac{w_{t-1}^{(i)}\widetilde{p}_t(X_t^{(i)})/\widetilde{p}_{t-1}(X_t^{(i)})}{\sum_{j=1}^N w_{t-1}^{(j)}\widetilde{p}_t(X_t^{(j)})/\widetilde{p}_{t-1}(X_t^{(j)})}; \tag{3.44}$$

but the resampling step, though it increases the variance locally at a given step t, is shown to help prevent degenerating importance weights. It is also possible to rearrange the steps and proceed in the order of (i) propagating, (ii) reweighting, and (iii) resampling, as in (17), but in this case the particles are instead propagated according to a $\mathcal{P}_{t-1}$-invariant Markov transition kernel

By using such a sequential approach, we could re-use the current samples from the PPD $p(x_0|y_0, \mathcal{D}_{t-1})$ (or from the joint $p(x_0, \theta|y_0, \mathcal{D}_{t-1})$) to obtain new samples once the dataset is augmented into $\mathcal{D}_t$.

- Could such SMC samplers, or perhaps more sophisticated variants, be applied in this context of sampling from the posterior predictive in combination with the Gibbs sampling scheme that we proposed for generative or discriminative PPD sampling?

- In particular, could such sequential algorithms withstand a high number of parameters, especially in NN-based models ?

Bibliography

[1] Alexander C Aitken. On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, 55:42–48, 1936.

[2] Ahmed Alaa and Mihaela Van Der Schaar. Discriminative jackknife: Quantifying uncertainty in deep learning via higher-order influence functions. In *International Conference on Machine Learning*, pages 165–174. PMLR, 2020.

[3] Elouan Argouarc'h, François Desbouvries, Eric Barat, Eiji Kawasaki, and Thomas Dautremer. Discretely Indexed Flows, 2022.

[4] Andrei Atanov, Alexandra Volokhova, Arsenii Ashukha, Ivan Sosnovik, and Dmitry Vetrov. Semi-conditional normalizing flows for semi-supervised learning. *arXiv preprint arXiv:1905.00505*, 2019.

[5] Atilim Gunes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of machine learning research*, 18(153):1–43, 2018.

[6] Joseph Berkson. Application of the logistic function to bio-assay. *Journal of the American statistical association*, 39(227):357–365, 1944.

[7] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[8] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[9] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

[10] Olivier Cappé, Simon J Godsill, and Eric Moulines. An overview of existing methods and recent advances in sequential Monte Carlo. *Proceedings of the IEEE*, 95(5):899–924, 2007.

[11] George Casella and Edward I George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

[12] Changyou Chen, David Carlson, Zhe Gan, Chunyuan Li, and Lawrence Carin. Bridging the gap between stochastic gradient MCMC and stochastic optimization. In *Artificial Intelligence and Statistics*, pages 1051–1060. PMLR, 2016.

[13] Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335, 1995.

[14] Nicolas Chopin, Omiros Papaspiliopoulos, et al. *An introduction to sequential Monte Carlo*, volume 4. Springer, 2020.

[15] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

[16] Katalin Csilléry, Michael GB Blum, Oscar E Gaggiotti, and Olivier François. Approximate Bayesian computation (ABC) in practice. *Trends in ecology & evolution*, 25(7):410–418, 2010.

[17] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(3):411–436, 2006.

[18] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

[19] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? Does it matter? *Structural safety*, 31(2):105–112, 2009.

[20] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[21] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[22] Conor Durkan, George Papamakarios, and Iain Murray. Sequential neural methods for likelihood-free inference. *arXiv preprint arXiv:1811.08723*, 2018.

[23] Bradley Efron. Jackknife-after-bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 54(1):83–111, 1992.

[24] Ethan Fetaya, Jörn-Henrik Jacobsen, Will Grathwohl, and Richard Zemel. Understanding the limitations of conditional generative models. *arXiv preprint arXiv:1906.01171*, 2019.

[25] Felix Fiedler and Sergio Lucia. Improved uncertainty quantification for neural networks with Bayesian last layer. *IEEE Access*, 2023.

[26] Edwin Fong, Chris Holmes, and Stephen G Walker. Martingale posterior distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5):1357–1391, 02 2024.

[27] Edwin Fong, Simon Lyddon, and Chris Holmes. Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap. In *International Conference on Machine Learning*, pages 1952–1962. PMLR, 2019.

[28] Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

[29] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[30] A. E Gelfand. Gibbs sampling. *Journal of the American statistical Association*, 95(452):1300–1304, 2000.

[31] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

[32] Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459, 2015.

[33] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[34] N.J. Gordon, D.J. Salmond, and A.F.M. Smith Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, 140:107–113(6), April 1993.

[35] Alex Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011.

[36] David Greenberg, Marcel Nonnenmacher, and Jakob Macke. Automatic posterior transformation for likelihood-free inference. In *International Conference on Machine Learning*, pages 2404–2414. PMLR, 2019.

[37] Lauren A Hannah, David M Blei, and Warren B Powell. Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, 12(6), 2011.

[38] Trevor Hastie, Robert Tibshirani, Jerome Friedman, Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Unsupervised learning. *The elements of statistical learning: Data mining, inference, and prediction*, pages 485–585, 2009.

[39] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[40] Geoffrey Hinton and Terrence J Sejnowski. *Unsupervised learning: foundations of neural computation*. MIT press, 1999.

[41] Chris Holmes and Leonhard Knorr-Held. Bayesian Auxiliary Variable Models for Binary and Multinomial Regression. *Bayesian Analysis*, 1:145–168, 03 2006.

[42] Ziyi Huang, Henry Lam, and Haofeng Zhang. Quantifying epistemic uncertainty in deep learning. *arXiv preprint arXiv:2110.12122*, 2021.

[43] Ziyi Huang, Henry Lam, and Haofeng Zhang. Efficient uncertainty quantification and reduction for over-parameterized neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.

[44] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

[45] Pavel Izmailov, Polina Kirichenko, Marc Finzi, and Andrew Gordon Wilson. Semi-supervised learning with normalizing flows. In *International Conference on Machine Learning*, pages 4615–4630. PMLR, 2020.

[46] Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Gordon Wilson. What are Bayesian neural network posteriors really like? In *International conference on machine learning*, pages 4629–4640. PMLR, 2021.

[47] Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. Hands-on Bayesian neural networks—A tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48, 2022.

[48] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[49] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[50] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. pages 32–33, 2009.

[51] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[52] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[53] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12):6999–7019, 2021.

[54] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461, 2003.

[55] Stuart Lloyd. Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[56] Jan-Matthis Lueckmann, Giacomo Bassetto, Theofanis Karaletsos, and Jakob H Macke. Likelihood-free inference with emulator networks. In *Symposium on Advances in Approximate Bayesian Inference*, pages 32–53. PMLR, 2019.

[57] Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke. Benchmarking simulation-based inference. In *International conference on artificial intelligence and statistics*, pages 343–351. PMLR, 2021.

[58] Simon Lyddon, Stephen Walker, and Chris C Holmes. Nonparametric learning from Bayesian models with randomized objective functions. *Advances in neural information processing systems*, 31, 2018.

[59] David JC MacKay. A practical Bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

[60] Radek Mackowiak, Lynton Ardizzone, Ullrich Kothe, and Carsten Rother. Generative classifiers as a basis for trustworthy image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2971–2981, 2021.

[61] James N Morgan and John A Sonquist. Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, 58(302):415–434, 1963.

[62] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[63] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[64] Michael A Newton, Nicholas G Polson, and Jianeng Xu. Weighted Bayesian bootstrap for scalable posterior distributions. *Canadian Journal of Statistics*, 49(2):421–437, 2021.

[65] Michael A Newton and Adrian E Raftery. Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 56(1):3–26, 1994.

[66] Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14, 2001.

[67] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

[68] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine learning*, 39:103–134, 2000.

[69] Augustus Odena. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*, 2016.

[70] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

[71] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30, 2017.

[72] George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *The 22nd international conference on artificial intelligence and statistics*, pages 837–848. PMLR, 2019.

[73] Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

[74] Juho Piironen and Aki Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11, 07 2017.

[75] Adrian E Raftery, David Madigan, and Jennifer A Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191, 1997.

[76] Michael Revow, Christopher KI Williams, and Geoffrey E Hinton. Using generative models for handwritten digit recognition. *IEEE transactions on pattern analysis and machine intelligence*, 18(6):592–606, 1996.

[77] Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *6th international conference on learning representations, ICLR 2018-conference track proceedings*, volume 6. International Conference on Representation Learning, 2018.

[78] Christian P Robert et al. *The Bayesian choice: from decision-theoretic foundations to computational implementation*, volume 2. Springer, 2007.

[79] G. O. Roberts and J.S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20 – 71, 2004.

[80] Tim GJ Rudner, Zonghao Chen, Yee Whye Teh, and Yarin Gal. Tractable function-space variational inference in bayesian neural networks. *Advances in Neural Information Processing Systems*, 35:22686–22698, 2022.

[81] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[82] Harald Steck and Tommi Jaakkola. On the Dirichlet prior and Bayesian regularization. *Advances in neural information processing systems*, 15, 2002.

[83] Ilkay Ulusoy and Christopher M Bishop. Generative versus discriminative methods for object recognition. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 258–265. IEEE, 2005.

[84] Matias Valdenegro-Toro and Daniel Saromo Mori. A deeper look into aleatoric and epistemic uncertainty disentanglement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1508–1516. IEEE, 2022.

[85] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[86] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

[87] Florian Wenzel, Kevin Roth, Bastiaan S Veeling, Jakub Swikatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the Bayes posterior in deep neural networks really? *arXiv preprint arXiv:2002.02405*, 2020.

[88] Peter M Williams. Bayesian regularization and pruning using a Laplace prior. *Neural computation*, 7(1):117–143, 1995.

[89] Chenyu Zheng, Guoqiang Wu, Fan Bao, Yue Cao, Chongxuan Li, and Jun Zhu. Revisiting discriminative vs. generative classifiers: Theory and implications. In *International Conference on Machine Learning*, pages 42420–42477. PMLR, 2023.

[90] Roland S Zimmermann, Lukas Schott, Yang Song, Benjamin Adric Dunn, and David A Klindt. Score-based generative classifiers. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021.

Chapter 4

Modeling a tractable PDF with DIF

The probability density function (PDF) plays a central role in statistics. It enables the computation of key quantities such as probabilities and moments of a given random variable (RV). In many hypothesis testing methods, the PDF is crucial for determining the likelihood of observed data under the null hypothesis, thus enabling us to conclude on the relevance of this hypothesis. Additionally, the PDF is fundamental to many learning tasks, facilitating parameter estimation through maximum likelihood or maximum a posteriori methods. More generally, PDFs are often involved in computing, estimating, or minimizing discrepancies between probability distributions. Finally, as we have explained in Chapter 2, PDFs are often used when it comes to sampling from the corresponding probability distribution with methods such as accept-reject, importance sampling, Markov chain Monte Carlo or Variational Inference (VI); and conversely, many MC estimation tasks aim at sampling from a distribution which is known via its unnormalized PDF.

In chapter 3, we highlighted the importance of epistemic uncertainty quantification and proceeded to a comparison of generative and discriminative modeling methods in Bayesian uncertainty-aware inference. These two modeling methods share a common trait: they use a parametric conditional model. In this work, we explained that both these modeling techniques are indeed compatible with the task of sampling from PPD if the corresponding model benefits from tractable PDF. However, in this context, a parametric probability distribution model must be appropriately constructed for fast and exact PDF evaluation, which is now the precise focus of this chapter.

This chapter is centered around the work on *Discretely Indexed Flows* (DIF) which corresponds to section 4.3. However, we first provide more insight into the context of this work. In section 4.1, we review existing generative modeling methods, sometimes referred to as density estimation in the literature, although the underlying model does not necessarily benefit from density evaluation. As we have mentioned in the previous chapter, generative modeling can refer to several notions. In Chapter 3, generative modeling referred to the task of approximating a posterior distribution via a parameterization of the conditional distribution over observations given labels. In the context of this chapter, generative modeling refers to the task of fitting and sampling from a probability distribution that closely resembles the one that produced some recorded data.

The main goal of DIF is to propose a generative modeling construction which indeed benefits from a tractable PDF, thus enabling density estimation. Therefore, we first review the usual methods for generative modeling and we particularly of emphasize the mechanisms that enable a model to benefit from a tractable PDF. Moreover, in the section covering DIFs, we contrast the problem of density estimation with the problem of VI. Therefore, for context, we introduce the related problem of gradient reparameterization in section 4.2 and describe the usual methods for estimating gradients of expectations computed with respect to parameterized probability distributions.

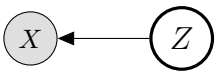
We then reach the culmination of this chapter in section 4.3, where we introduce and motivate DIF, a methodology for constructing parametric probability distributions that benefit from (i) fast and exact evaluation of their density function, (ii) improved flexibility in variational estimation problems using neural-network (NN) functions, and (iii) the ability to be easily turned into a conditional model, and thus to be used in a generative or discriminative posterior modeling in the context of the previous chapter.

4.1 Generative modeling and tractable PDF

In this section, we briefly review existing methods for generative modeling. They take interest in building a probability distribution that is, by construction, easy to sample from and which we can fit to recorded observations $x_1, ..., x_N \overset{iid}{\sim} \mathcal{P}$. In this context, we specifically focus on the mechanisms for building parametric probability distributions which benefit from tractable PDF, as we will leverage these mechanisms when introducing the DIF construction.

4.1.1 Latent variable models

A latent variable model (LVM) is a probabilistic model in which a RV of interest (referred to as *observed*) has a probability distribution which is related to one (or several) unobserved, *latent* RV via a conditional probability distribution. More precisely, we denote Z a hidden RV distributed according to $\mathcal{P}(Z)$ and we denote X an observed RV related to Z via a conditional probability distribution $\mathcal{P}(X|Z)$. The distribution over RV X can therefore be described as a directed graph and can be sampled from using the sequential sampling procedure: $z \sim \mathcal{P}(Z)$ and $x \sim \mathcal{P}(X|Z = z)$. We provide the corresponding directed graph:



$$\begin{aligned} Z &\sim \mathcal{P}(Z) \\ X &\sim \mathcal{P}(X|Z) \end{aligned} \tag{4.1}$$

As a result, on the one hand, the joint PDF of the two RVs can be written as $p(x, z) = p(z)p(x|z)$ and on the other hand, the PDF of the distribution over the observed variable is the x-marginal in this joint PDF.

In practice, LVMs can be used for generative modeling by considering a joint PDF $p_{\theta,\text{LVM}}(x, z)$ parameterized by θ and that we aim at adjusting θ so that the corresponding model fits the recorded $x_1, ..., x_N$. A possible approach, which we consider for illustration

purposes, is that of *Maximum-Likelihood Estimation* (MLE), where we want to obtain a value θ^* which reads:

$$\theta^* \in \arg\max_{\theta\in\Theta} \underbrace{\sum_{i=1}^{N} \log\big(p_{\theta,\mathrm{LVM}}(x_i)\big)}_{f(\theta)}. \tag{4.2}$$

However in general, the PDF of interest $p_{\theta,\mathrm{LVM}}(x) = \int p_{\theta,\mathrm{LVM}}(x,z)\mathrm{d}z$ can not be evaluated since the integral does not admit a closed form expression which raises the question of solving the MLE problem. In this context, we start off by explaining the principles of MLE for LVMs in section 4.1.1 and illustrate the application of this principle to Variational AutoEncoder (VAE) in section 4.1.1 using the Evidence Lower BOund (ELBO) as a Minorize-Maximize (MM) scheme. We then come to LVMs with a tractable PDF with mixture models in section 4.1.1 and present the corresponding Expectation-Maximization (EM) algorithm.

MLE for LVMs using MM, ELBO and EM

In this section, we denote $f(\theta)$ the function to be optimized, see (4.2). When considering LVMs, two cases arise. On the one hand, in general, the function $f(\theta)$ to be optimized is intractable, in which case the ELBO principle leverages the latent variable structure to obtain an alternative yet coherent optimization objective. On the other hand, when using specific LVM constructions, the function $f(\theta)$ can indeed be tractable (as is the case when the latent RV is categorical, see section 4.1.1) and can be optimized using standard (possibly gradient-based) methods. In this case, the EM algorithm can nonetheless be applied to produce efficient parameter updates, leading to an overall efficient optimization algorithm.

The EM algorithm (24) as well as variational EM with the ELBO function (50)(18) can both be understood as specific applications of the MM (43)(82) algorithm approach, which leverages (i) the latent structure of the underlying model and (ii) Jensen's inequality to construct an optimization surrogate function. We now start off by describing the MM principle.

The MM approach builds a series of surrogate functions $g_t(\theta)$ in a two-step sequential procedure. From the current value $\theta^{(t)}$, we first obtain a function $g_t(\theta)$ which satisfies the two conditions:

$$g_t(\theta) \leq f(\theta) \tag{4.3}$$

$$g_t(\theta^{(t)}) = f(\theta^{(t)}). \tag{4.4}$$

Then, from the current surrogate function g_t, we will deduce a point $\theta^{(t+1)}$ which increases the value of the surrogate:

$$g_t(\theta^{(t+1)}) \geq g_t(\theta^{(t)}). \tag{4.5}$$

By doing so, we ensure that the values $\{f(\theta)\}_t$ increase to a local maximum of f since:

$$f(\theta^{(t+1)}) \overset{(4.3)}{\geq} g_t(\theta^{(t+1)}) \overset{(4.5)}{\geq} g_t(\theta^{(t)}) \overset{(4.4)}{=} f(\theta^{(t)}).$$

This last step can be conducted by explicit maximization of the surrogate function, or by a Gradient Ascent (GA) step (this is discussed in the next sections).

We now apply the principle of MM to LVMs in order to obtain the ELBO objective. To this end, we will construct such a surrogate for each of the functions $\log(p_{\theta,\text{LVM}}(x_i))$ and deduce a surrogate for $f(\theta) = \sum_{i=1}^{M} \log(p_{\theta,\text{LVM}}(x_i))$ as the sum of the individual surrogates. Therefore, let us consider $\mathcal{L}_i(Z)$ an arbitrary distribution over the latent variable Z, and let us denote $l_i(z)$ its PDF. This enables us to rewrite $\log(p_{\theta,\text{LVM}}(x_i))$ as follows:

$$\log\big(p_{\theta,\text{LVM}}(x_i)\big) = \int \log\left(\frac{p_{\theta,\text{LVM}}(x_i,z)}{l_i(z)}\right) l_i(z)\mathrm{d}z + D_{\text{KL}}\big(l_i(z)||p_{\theta,\text{LVM}}(z|x_i)\big). \tag{4.6}$$

Since the Kullback-Leibler Divergence (D_{KL}) is always positive, we have that $\log(p_\theta(x_i)) \geq \int \log(\frac{p_\theta(x_i,z)}{l_i(z)}) l_i(z)\mathrm{d}z$ with equality if and only if $l_i(z) = p_\theta(z|x_i)$ almost surely (in which case the D_{KL} is zero). On the one hand, if we sum over values $x_1, ..., x_M$, we obtain the ELBO, which reads:

$$\text{ELBO}(\theta, l_1, ..., l_M) = \sum_{i=1}^{M} \mathbb{E}_{Z\sim\mathcal{L}_i}\left[\log\left(\frac{p_\theta(x_i,Z)}{l_i(Z)}\right)\right]; \tag{4.7}$$

which, as its name suggests, is, by construction, a lower bound for the likelihood, and the closer $l_i(z)$ is to $p_\theta(z|x_i)$, the tighter this bound is (see (72) for a discussion).

The EM algorithm builds upon the principle of the ELBO to obtain an exact MM scheme for maximizing the function of interest $f(\theta)$, where the ELBO is used as an optimization surrogate function. Indeed, from the ELBO (4.7), if the posterior probability distribution is tractable, one can obtain an MM scheme by setting $l_i(z)$ to be the current posterior distribution $\mathcal{P}_{\theta^{(t)}}(Z|x_i)$ which yields a surrogate which reads:

$$g_t(\theta) = \sum_{i=1}^{M} \mathbb{E}_{Z\sim\mathcal{P}_{\theta^{(t)}}(Z|x_i)}\left[\log\left(\frac{p_\theta(x_i,Z)}{p_{\theta^{(t)}}(Z|x_i)}\right)\right]; \tag{4.8}$$

and which satisfies, by construction, the two conditions of the MM principle (4.3) and (4.4). This first step of constructing $g_t(\theta)$ from the current value of θ is classically referred to as the E-step since it involves an expectation over the current latent variable posterior distribution (hence the denomination *Expectation*-Maximization). It now remains to find a point $\theta^{(t+1)}$ such that $g_t(\theta^{(t+1)}) \geq g_t(\theta^{(t)})$. This can be achieved by GA on the function $g_t(\theta)$ or by explicit maximization with $\theta^{(t+1)} \in \arg\max_{\theta\in\Theta} g_t(\theta)$. This step is classically referred to as the M-step since we aim to increase the surrogate and ideally maximize it (hence the denomination Expectation-*Maximization*). In the next section covering mixture models, we will obtain an update of parameter θ explicitly by solving $\nabla_\theta g_t(\theta)|_{\theta=\theta^{(t)}} = 0$. In this section, we specifically examine the case where the M-step is performed via GA:

$$\theta^{(t+1)} = \theta^{(t)} + \eta_t \nabla_\theta g_t(\theta)|_{\theta=\theta^{(t)}}, \tag{4.9}$$

for sufficiently small values of η_t, this parameter update on θ verifies the third condition of the MM algorithm. We now explain that this procedure is indeed equivalent to

performing GA on $f(\theta)$ since the value of the gradient $\nabla_\theta g_t(\theta)|_{\theta=\theta^{(t)}}$ is the same as $\nabla_\theta f(\theta)|_{\theta=\theta^{(t)}}$. It is indeed clear from the MM construction that the function $h : \theta \longrightarrow g_t(\theta) - f(\theta)$ is negative (see (4.3)) and is 0 for $\theta = \theta^{(t)}$ (see (4.4)). So the point $\theta^{(t)}$ is a maximum of h and $\nabla_\theta h(\theta)|_{\theta=\theta^{(t)}} = 0$, which finally yields:

$$\nabla_\theta g_t(\theta)|_{\theta=\theta^{(t)}} - \nabla_\theta f(\theta)|_{\theta=\theta^{(t)}} = 0. \tag{4.10}$$

Even though the gradient based EM is equivalent to performing the standard GA of f, this gradient based EM algorithm can nonetheless be relevant in cases where either $f(\theta)$ or its gradient can not be computed while the gradient of $g_t(\theta)$ can be computed. In the context of mixture models, we will illustrate how the ELBO surrogate admits closed-form M-step updates. The EM algorithm is a very efficient and well-studied (87)(10) parameter estimation technique which has become a very popular tool to solve variational density estimation (VDE) tasks (64) and which has been applied in a wide range of contexts such as signal processing (23) and image reconstruction (31). In particular, when applied to the GMM which are multimodal by nature, they are often used for fast and accurate unsupervised clustering (17).

However, the EM algorithm that we have just presented, be it with explicit surrogate maximization or via gradient updates, (and for which we will describe its application to mixture models in section 4.3.10), requires (i) computing the posterior PDF of latent variables $p_{\theta^{(t)}}(Z|x_i)$ and (ii) computing expectation with respect to this posterior distribution, see (4.8). However, when these two conditions are not fulfilled, one can resort to the variational expectation maximization of the ELBO. This principle is used in VAEs, which is described in the next section 4.1.1.

VAE

A VAE (50) is a specific LVM where (i) the latent variable is distributed according to a parameter-free distribution, such as a standard normal, and (ii), given Z, X is distributed according to a conditional distribution parameterized by θ, for example, $p_\theta(x|z) = \mathcal{N}(\mu_\theta(Z), \sigma^2_\theta(Z))$ where $\mu_\theta(Z)$ and $\sigma^2_\theta(Z)$ are the output of a NN function parameterized by θ. The corresponding PDF reads:

$$p_{\theta,\text{VAE}}(x) = \int p_\theta(x|z)q(z)\mathrm{d}z. \tag{4.11}$$

For arbitrary functions μ_θ and σ_θ, this PDF does not admit a closed form expression which calls for the ELBO (4.7). However, in this setting, the exact EM algorithm is not applicable since the posterior PDF $p_\theta(z|x)$ cannot be expressed in closed form and can only be evaluated up to a normalizing constant. Therefore, in the seminal paper, the author considers the ELBO optimization objective in which the distributions $l_i(Z)$ are defined via a variational approximation of the posterior probability distribution, which also uses NN functions (say parameterized by ϕ) μ_ϕ and σ^2_ϕ such that $l_i(z) \triangleq p_\phi(z|x_i) = \mathcal{N}(z; \mu_\phi(x_i), \sigma^2_\phi(x_i))$. The ELBO then becomes:

$$\text{ELBO}(\theta, \phi) = \sum_{i=1}^{M} \mathbb{E}_{Z\sim\mathcal{P}_\phi(Z|X=x_i)}\left[\log\left(\frac{p_\theta(x_i, Z)}{p_\phi(Z|x_i)}\right)\right]. \tag{4.12}$$

An unbiased Monte-Carlo estimate of the gradient of this expression with respect to both parameters θ and ϕ (12) is then computed using the reparameterization gradient method (86) or the reparameterization trick (53). There are many further constructions of LVMs which extend the principle of VAEs, such as deep and stacked variational autoencoding layers (81), hierarchical VAEs (85), or continuously indexed NFs (20).

Mixture Model

As we have mentioned before, the PDF associated with a LVM $p_{\theta,\text{LVM}}(x)$ cannot be computed in general because the integral $\int p_{\theta,\text{LVM}}(x,z)\text{d}z$ does not admit a closed-form expression. However, if the latent RV is discrete with finite values, then the integral indeed becomes a finite sum and is tractable. A possible approach to constructing a probability distribution with tractable PDF is therefore to consider a categorical latent variable, which we now denote R, that takes values $r = 1, ..., K$ with probability $\pi_k \triangleq \Pr(R = k)$. The PDF associated with RV R is therefore a probability mass function which reads $\sum_{k=1}^{K} \pi_k \delta_{r,k}$, where $\delta_{.,.}$ is the delta Kronecker. Let us denote $\mathcal{P}_k$ the distribution of the observed RV X given $R = k$. The corresponding construction can be viewed as considering $K + 1$ RVs: $X_k \sim \mathcal{P}_k$ for $k = 1, ..., K$ and $R \sim$ Categorical$\{\pi_1, ..., \pi_K\}$; and defining the RV $X = \sum_{k=1}^{K} X_k \delta_{R,k} = X_R$. Then this RV is said to be distributed according to the mixture of distributions $\mathcal{P}_1, ..., \mathcal{P}_K$ with weights $\pi_1, ..., \pi_K$. The PDF associated with the corresponding model finally reads:

$$p_{\text{Mixture}}(x) = \sum_{k=1}^{K} \pi_k p_k(x). \tag{4.13}$$

To summarize, a mixture model is a probability distribution obtained using two principles. As we explained, the first principle is that of a discrete finite latent variable, which ensures tractability during integration of the latent RV, and the second is that the latent variable has a distribution which is parameterized independently of the value of X to scalar values π_k. In the rest of this chapter, we will reach the DIF construction, which also leverages the principle of a discrete finite latent variable, but where the parameterization of the distribution associated with the categorical RV indeed depends on the value of X (see sections 4.1.3 and 4.3).

We now consider the previous mixture model and see how the EM algorithm for MLE is applied. We therefore consider a mixture of distributions $\mathcal{P}_k$, each parameterized by λ_k, and the mixture weights π_k. The parameters of the corresponding model are $\theta = \left[\pi_1, ..., \pi_K, \lambda_1, ..., \lambda_K\right]$. We first explicit the E-step in which we compute, from the current parameter value $\theta^{(t)}$, the surrogate function $g_t(\theta)$. The posterior probability that the latent variable R takes the value k for observation x_i, $\Pr_{\theta^{(t)}}(R = k|X = x_i)$, and which we denote as $v_k^{(t)}(x_i)$, reads:

$$v_k^{(t)}(x_i) \triangleq \Pr_{\theta^{(t)}}(R = k|X = x_i) = \frac{\pi_k^{(t)} p_k(x_i; \lambda_k^{(t)})}{\sum_{l=1}^{K} \pi_l^{(t)} p_l(x_i; \lambda_l^{(t)})}. \tag{4.14}$$

Since the latent variable is categorical (i.e. discrete with finite values), the expectations computed with respect to the posterior distributions are expressed as a finite sum, which

enables us to deduce (and compute) the surrogate function:

$$g_t(\theta) = \sum_{i=1}^{M} \sum_{k=1}^{K} v_k^{(t)}(x_i) \log\left(\frac{\pi_k p_k(x_i;\lambda_k)}{v_k^{(t)}(x_i)}\right). \tag{4.15}$$

So in the case of such mixture models, the E-step can be conducted explicitly.

We now come to the M-step, which consists in updating the parameters by increasing the surrogate computed in the previous E-step. In general, LVMs do not admit explicit maxima, but in that case, as we mentioned, it is only required that we increase the surrogate (possibly via a gradient-based method). However, in the specific case of mixture models, it turns out that we can maximize the mixture weights explicitly, so we now first derive the update for the mixture weights. First, we see in equation (4.15) that the parameters π_k are maximized independently of λ_l, $l = 1, ...,$ since the partial derivative of g_t with respect to π does not depend on the value of λ_l. At first glance, we might similarly conclude that π_k is maximized independently of the other values π_l, $l = 1, ..., K$ but remember that $\pi_1, ..., \pi_K$ are the probabilities of the categorical distribution of the latent variable R so they must sum to 1. We therefore first maximize equation (4.15) with respect to weight parameter π_k under the constraint that $\sum_{l=1}^{K} \pi_l = 1$ (or equivalently that the function $\pi_k \longrightarrow \left(\sum_{l=1}^{K} \pi_l\right) - 1$ is 0). To that end, we use the Lagrangian function with multiplier γ:

$$L(\pi_k, \gamma) = \log(\pi_k)\left(\sum_{i=1}^{M} v_k^{(t)}(x_i)\right) + \gamma\left(\sum_{l=1}^{K} \pi_l - 1\right); \tag{4.16}$$

and we aim to find the stationary points of L, i.e. to solve for $\pi_k^{(t+1)}, \gamma^*$ in $\nabla L(\pi_k^{(t+1)}, \gamma^*) = 0$. This leads a system of two equations, and its solution for $\pi_k^{(t+1)}$ reads:

$$\pi_k^{(t+1)} = \frac{\sum_{i=1}^{M} v_k^{(t)}(x_i)}{\sum_{l=1}^{K}\sum_{i=1}^{M} v_l^{(t)}(x_i)}. \tag{4.17}$$

We now consider maximizing the surrogate function with respect to the parameter λ_k. We see again in equation (4.15) that the parameter λ_k is maximized independently of π_l and other parameters λ_l, $l = 1, ..., K$ since again the partial derivative (or gradient if λ_k is multidimensional) does not depend on neither π_l nor λ_l. The function to maximize therefore reads:

$$L(\lambda_k) = \sum_{i=1}^{M} \sum_{k=1}^{K} v_k^{(t)}(x_i) \log\big(p_k(x_i;\lambda_k)\big). \tag{4.18}$$

We therefore seek λ_k^* which maximize this function by solving $\nabla L(\lambda_k^*) = 0$. Unlike for the mixture weights, a closed form update for the parameters of the distributions $\mathcal{P}_k$ is not necessarily available. Nonetheless, many different probability distributions indeed admit closed form updates, such as Gaussian (24) or Poisson (11) distributions in which case the EM algorithm is straightforward. We now consider the case of a Gaussian distribution so $p_k(x;\lambda_k) = \mathcal{N}(x;\mu_k,\Sigma_k)$ where $\lambda_k = \big\{\mu_k, \Sigma_K\big\}$ and μ_k is the mean vector and Σ_k is a covariance (symmetric positive-definite) matrix. The corresponding

mixture model is therefore a Gaussian Mixture Model (GMM). We recall that $\mathcal{N}$ is a notation used to describe the PDF associated with a Gaussian distribution:

$$\mathcal{N}(x;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right). \tag{4.19}$$

The function to maximize with respect to μ_k and Σ_k (after removing constant terms) therefore reads:

$$L(\mu_k,\Sigma_k) = \sum_{i=1}^{M} \frac{v_k^{(t)}(x_i)}{2}\Big(\log\big(\det(\Sigma_k^{-1})\big) - \big(x_i-\mu_k\big)^{\mathrm{T}}\Sigma_k^{-1}\big(x_i-\mu_k\big)\Big). \tag{4.20}$$

Instead of computing the partial derivative of L with respect to Σ_k, we instead compute the partial derivative of $L'(\mu_k,\Sigma_k^{-1}) \triangleq L(\mu_k,\Sigma_k)$ with respect to Σ_k^{-1} since ultimately $\nabla L'(\mu_k^{(t+1)},{\Sigma_k^{(t+1)}}^{-1}) = 0 \iff L(\mu_k^{(t+1)},\Sigma_k^{(t+1)}) = 0$. One can compute the derivative with respect to Σ_k^{-1} with matrix derivatives identities, and use the properties arising from Σ_k being a covariance matrix to solve for μ_k^*,Σ_k^* in $\nabla L'(\mu_k^{(t+1)},{\Sigma_k^{(t+1)}}^{-1}) = 0$. This yields a system of equation and leads to the solution of the parameter update for mean and covariance matrix during the M-step in a GMM:

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{M} v_k^{(t)}(x_i)x_i}{\sum_{i=1}^{M} v_k^{(t)}(x_i)} \tag{4.21}$$

$$\Sigma_k^{(t+1)} = \frac{\sum_{i=1}^{M} v_k^{(t)}(x_i)\Big(x_i-\mu_k^{(t+1)}\Big)\Big(x_i-\mu_k^{(t+1)}\Big)^{\mathrm{T}}}{\sum_{i=1}^{M} v_k^{(t)}(x_i)} \tag{4.22}$$

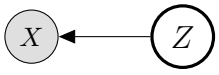
So, in the specific case of GMM, the EM algorithm can be conducted explicitly. Indeed, since it is a *mixture model*, it follows that (i) the E-step is straightforward as it is nothing but computing expectation with respect to a categorical distribution, and (ii) the mixture weights admit closed-form updates in the M-step. Moreover, since the mixture distributions are Gaussian, the parameters of the corresponding distribution can also be updated in closed form during the M-step.

4.1.2 Push-forward models

As we have explained in the previous section, LVMs indeed define an underlying directed graph in which the transition between a latent variable and a visible variable is defined via a parameterized probability distribution. In that regard, a concurrent approach to obtaining a parameterized probability distribution is also to consider a latent variable with some random probability distribution and relate it to an observed RV via a deterministic function, referred to as a push-forward function in this context. We now briefly describe the example of Generative Adversarial Networks (GANs), which are precisely built using this construction, and then we come to Normalizing Flows (NFs) which are also built using the push-forward principle but benefit from tractable PDF with the mechanism of change of variable, which we also describe.

GAN

GAN (37) are a class of parametric probability distribution in which a RV of interest is built as $T_\theta(Z)$ where Z is a random "noise" variable distributed according to some, usually parameter-free, distribution $\mathcal{P}(Z)$, and T_θ is an arbitrary NN function parameterized by θ.



$$\begin{aligned} Z &\sim \mathcal{P}(Z) \\ X &= T_\theta(Z) \end{aligned} \tag{4.23}$$

It is therefore easy to draw the corresponding RV by first sampling the noise value Z and passing it through T_θ. However, the PDF of the corresponding distribution $\mathcal{P}_{\theta,\text{GAN}}$ reads:

$$p_{\theta,\text{GAN}}(x) = \int \delta_{T_\theta(x)}(z) q(z) \mathrm{d}z, \tag{4.24}$$

and, for an arbitrary function T_θ, is intractable in general as the integral cannot be computed in closed form. This therefore raises the question of fitting such a model to recorded data, as standard MLE is unfeasible without explicit access to the PDF. The solution proposed in the seminal paper (37) is adversarial training. The concept of adversarial training leverages a classifier r_ϕ, which aims to distinguish between the recorded samples $x_1, ..., x_M$ and samples from $\mathcal{P}_{\theta,\text{GAN}}$. Intuitively, the goal is to adjust θ such that the samples from the model are indistinguishable from the distribution that produced the recorded data, in which case the two distributions are close to being the same. More formally, the considered optimization objective is:

$$L(\theta, \phi) = -\sum_{i=1}^{M} \log(r_\phi(x_i)) - \sum_{i=1}^{N} \log(1 - r_\phi(T_\theta(z_i)), \text{ where } z_1, ...z_N \sim \mathcal{P}(Z). \tag{4.25}$$

This expression is minimized with respect to ϕ and, for fixed θ corresponds to the usual Binary Cross Entropy criterion to build a classifier which distinguishes between recorded samples and samples from the corresponding GAN model; and this expression is maximized with respect to θ for fixed ϕ such that the corresponding model produces samples that can fool the classifier. The convergence of adversarial training has been thoroughly studied (40)(2)(60). Generative adversarial networks have constituted a major advance in generative modeling as, with leveraging deep NN functions for T_θ and r_ϕ, the corresponding model is able to produce highly accurate variational approximations of complex and high-dimensional probability distributions. GAN has made an impact in the specific context of generating natural images, and its mechanism and the principle of adversarial training have since been built upon to propose more elaborate models that indeed produce high-resolution and highly realistic images (47)(48).

NF

As we have seen in the previous section, parametric probability distributions can be defined as a push-forward transformation of a given base distribution. However, such models do not benefit from a tractable density function in general, and for estimating the model parameters, practitioners must then resort to more involved training procedures, such as adversarial training. Nonetheless, in the case where the push-forward

transformation is a change of variable, the density function indeed becomes tractable, which is the principle upon which the NF modeling technique, which we now describe, is based.

NFs are a class of parameterized probabilistic models which can be related to the principle of inverse transform sampling, which we first briefly explain. Let X be a $\mathcal{P}$-distributed continuous real-valued RV (we consider the multivariate case hereafter). We denote P the Cumulative Distribution Function (CDF): $P(x) = \Pr(X \leq x)$. It then follows that the RV $P(X)$ is distributed uniformly on $[0, 1]$. To see this, we suppose for simplicity that the restriction of $\mathcal{P}$ in the support of $\mathcal{P}$ has inverse P^{-1} (if it is not the case, one can use similar arguments with using the generalized inverse distribution function (46)). In this case, $\Pr(P(X) \leq x) = \Pr(X \leq P^{-1}(x)) = P(P^{-1}(x)) = x$; which is nothing but the CDF of a uniform RV on $[0, 1]$. Similarly, let $U \sim \mathcal{U}[0, 1]$ be a uniformly distributed RV. It follows, with a similar argument, that $P^{-1}(U)$ is distributed according to $\mathcal{P}$. Indeed: $\Pr(P^{-1}(U) \leq x) = \Pr(U \leq P(x)) = P(x)$. We can therefore deduce a two-step procedure for sampling from $\mathcal{P}$ using the inverse CDF P^{-1}: we first sample uniformly $u \sim \mathcal{U}[0, 1]$ and we then compute $x = P^{-1}(u)$, which is hence a sample from $\mathcal{P}$.

The previous argument can easily be adapted in the case where $\mathcal{P}$ is a multivariate distribution, say of dimension d, by considering the function $P : \mathbb{R}^d \to \mathbb{R}^d$ which reads:

$$P(x)_1 = \Pr(X_1 \leq x_1) \text{ and} \tag{4.26}$$

$$P(x)_i = \Pr(X_i \leq x_i | X_1 = x_1, ..., X_{i-1} = x_{i-1}) \text{ for } i = 2, ..., d; \tag{4.27}$$

(with slightly abusing notations, we also refer to this function as CDF in the rest of this chapter). Indeed, it then follows that (i) if $X \sim \mathcal{P}$, then $P(X) \sim \mathcal{U}_{[0,1]^d}$, (ii) this function is invertible and (iii) if $U \sim \mathcal{U}_{[0,1]^d}$, then $P^{-1}(U) \sim \mathcal{P}$. We refer to (67) and the references therein (notably (44)(9)) for a more in-depth treatment of this fact.

Under mild conditions, this function is also differentiable, so it is a change of variables. As a consequence, for any distribution of interest $\mathcal{P}$, there exists a change of variables that transforms it into a given base distribution, be it a product of independent uniform distributions or a isotropic multivariate normal distribution. an NF defines a probability distribution $\mathcal{P}_{\theta,\text{NF}}$ which is parameterized by θ, via a change of variables T_θ which is applied to a base reference distribution $\mathcal{Q}$ (usually chosen as a standard normal distribution) and parameters θ are adjusted such that T_θ approximates the corresponding change of variable. We describe how one can use NFs for generative modeling and VI in detail in section 4.3.3.

For now we emphasize that the corresponding probability distribution benefits from tractable density via the change of variables formula:

$$p_{\theta,\text{NF}}(x) = q_\theta(T_\theta(x)) |\det J_{T_\theta}(x)|. \tag{4.28}$$

Therefore, specific push-forward models which use a change of variables to transform a RV distributed according to a base distribution indeed benefit from a tractable PDF, and this construction precisely corresponds to NFs.

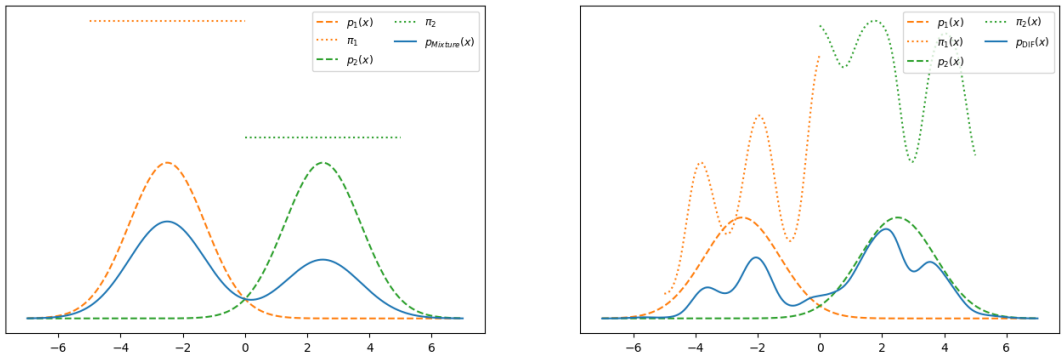


Figure 4.1: In the DIF construction, the constant mixture weights are replaced by a function

4.1.3 From Mixture Models Towards Discretely Indexed Flows

In the previous sections, we have explained the principle of mixture models and how one can build a tractable PDF with a linear combination of tractable PDFs. We then illustrated this principle with the GMM, which is considered by most practitioners as a very efficient tool for solving density estimation problems thanks to the potency of the EM optimization algorithm, which, as we have explained, admits closed form updates. Moreover, the GMM is a universal approximation tool for density functions (13) (71). However, the flexibility of a GMM, i.e., its ability to capture details and obtain fine approximations of a given distribution, is only related to its number of components; the higher the number of components, the higher the capacity of the GMM to closely approximate any density function. In this context, and in an attempt to induce further flexibility in a mixture model while retaining the benefit of a tractable PDF, we propose the construction of DIF. The idea of DIF is to increase the flexibility of mixture models with a fixed number of components by replacing the mixture weights $\pi_k \in [0, 1]$ by a flexible function $\pi_k(x)$.

More precisely, let us go back to the probability distribution which corresponds to the mixture of distributions $\mathcal{P}_k$ with weights $\pi_1, ..., \pi_K$. We aim to extend this mixture model into a more general construction where the mixture weights depend on the observed variable X. Our goal is to write a PDF of the form:

$$p_{\text{Mixture}}(x) = \sum_{k=1}^{K} \pi_k p_k(x) \Longrightarrow p_{\text{DIF}}(x) = \sum_{k=1}^{K} \pi_k(x) p_k(x). \tag{4.29}$$

The objective of the DIF construction is illustrated in figure 4.1.

In section 4.3, we cover in detail the DIF modeling technique and its construction is presented as an extension of NFs. Yet, to complement the understanding of this methodology, we first relate DIF to mixture models with regard to equation 4.29. The goal of this section is to unravel sufficient conditions on the functions $\pi_k(x)$ for which this equation indeed corresponds to a valid PDF. We point out that the construction of a DIF involves both the principles of a discrete latent variable (hence the term *Discretely*) and a change of variables from some base distribution (hence the term *Flow*).

The two properties for $p_{\text{DIF}}(x)$ to be a valid PDF are that the function is non-negative everywhere and that its integral is 1. So we aim at fulfilling the two requirements:

$$\sum_{k=1}^{K} \pi_k(x) p_k(x) \geq 0, \forall x \in \mathbb{R}^d; \tag{4.30}$$

$$\int \sum_{k=1}^{K} \pi_k(x) p_k(x) \mathrm{d}x = 1. \tag{4.31}$$

First, since $p_k(x)$ is a PDF, it is itself non-negative, and therefore a sufficient condition on the function π_k such that (4.30) is satisfied is that $\pi_k(x) \geq 0$ for all $k = 1, ..., K$ and for all $x \in \mathbb{R}^d$. We now discuss the second condition, and to that end, we will again use the CDF and the principle of inverse CDF transform sampling to find a sufficient condition on functions π_k such that (4.31) is satisfied. Using P_k the CDF of $\mathcal{P}_k$ and the change of variable $u = P_k(x)$, we rewrite the left-hand side of (4.31) as:

$$\int \sum_{k=1}^{K} \pi_k(x) p_k(x) \mathrm{d}x = \int_{[0,1]^d} \sum_{k=1}^{K} \pi_k(P_k^{-1}(u)) \mathrm{d}u. \tag{4.32}$$

We can deduce a sufficient condition about functions π_k to satisfy (4.31): we see that if $\sum_{k=1}^{K} \pi_k(P_k^{-1}(u)) = 1$ for all value $u \in [0, 1]$, then this integral indeed reduces to $\int_{[0,1]^d} \mathrm{d}u = 1$. We have hence identified two sufficient conditions on the functions π_k such that the induced model is indeed a valid probability distribution:

$$\sum_{k=1}^{K} \pi_k(P_k^{-1}(u)) = 1 \text{ and } \pi_k(x) \geq 0, \text{ for } x \in \mathbb{R}^d. \tag{4.33}$$

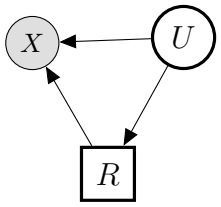
Up to this point, it is perhaps still unclear how to construct functions π_k such that these two requirements are satisfied. The final step is to rewrite $\alpha_k(u) \triangleq \pi_k(P_k^{-1}(u)) \iff \pi_k(x) = \alpha_k(P_k(x))$, and rewrite the two previous sufficient conditions on π_k in terms of functions α_k:

$$\sum_{k=1}^{K} \alpha_k(u) = 1 \text{ and } \alpha_k(u) \geq 0 \text{ for all } u \in [0, 1]^d \text{ and } k = 1, ..., K; \tag{4.34}$$

We see that the condition is that functions $\alpha_k(u)$ predict a vector of probabilities, hence, it defines a categorical probability distribution, and the value $\alpha_k(u)$ can therefore be interpreted as the probability that the uniform RV U is transformed into $P_k^{-1}(U)$ given that U takes the value u. So the random categorical latent variable R takes the value $k = 1, ..., K$ with probability:

$$\Pr(R = k | U = u) = \alpha_k(u) = \Pr(X = P_k^{-1}(U) | U = u). \tag{4.35}$$

The construction of the DIF therefore corresponds to an LVM with two latent variables: U distributed Uniformly and R a categorical RV with probabilities that depend on the value of U via functions $\alpha_1(U), ..., \alpha_K(U)$.

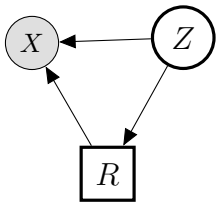


$$\begin{aligned} &U \sim \mathcal{U}[0, 1] \\ &R \sim \text{ Categorical } [\alpha_1(U), ..., \alpha_K(U)] \\ &X = P_R^{-1}(U) \end{aligned} \tag{4.36}$$

We have considered the inverse CDF transform P_k^{-1} of $\mathcal{P}_k$ which, as we explained before, is a change of variables between $\mathcal{P}_k$ and the uniform distribution $\mathcal{U}_{[0,1]^d}$. But again, we can instead consider a change of variables between $\mathcal{P}_k$ and any base distribution $\mathcal{Q}$ with tractable density by defining $T_k^{-1} = Q \circ P_k^{-1}$ where Q is the CDF of $\mathcal{Q}$. In this case, we define $w_k(z) \stackrel{\Delta}{=} \pi_k(T_k^{-1}(z)) = \pi_k(P_k^{-1}(Q(z)) = \alpha_k(Q(z))$; and the two conditions on functions α_k translate to conditions on functions w_k:

$$\sum_{k=1}^K w_k(z) = 1 \text{ and } w_k(z) \geq 0, \text{ for all } z \in \mathbb{R}^d \text{ and } k = 1, ..., K; \tag{4.37}$$

$$\sum_{k=1}^K w_k(z) = 1, \text{ for all } z \in \mathbb{R}^d. \tag{4.38}$$



$$\begin{aligned} Z &\sim \mathcal{Q} \\ R &\sim \text{ Categorical } [w_1(Z), ..., w_K(Z)] \\ X &= T_R^{-1}(U) \end{aligned} \tag{4.39}$$

We finally see that conditions on functions w_k are that they define a vector of categorical probability for all values $z \in \mathbb{R}^d$ and $w_k(z)$ can be considered as the output of a classifier function. We can also rewrite the density of the corresponding model as $p_{\text{DIF}}(x) = \sum_{k=1}^K w_k(T_k(x))p_k(x)$. Again, similarly to the principle of NF, we can rewrite the density associated with $p_k(x)$ in terms of q, the PDF associated with $\mathcal{Q}$, using the change of variables formula for densities, as $p_k(x) = q(T_k(x))|\det J_{T_k}(x)|$. Finally, the density of the corresponding DIF model reads:

$$p_{\text{DIF}}(x) = \sum_{k=1}^K w_k(T_k(x))q(T_k(x))|\det J_{T_k}(x)|. \tag{4.40}$$

So finally, we can summarize the principle of the DIF mechanism. We considered a mixture model where the mixture weights are replaced by a classifier function which computes the categorical probabilities for a standardized version of the variable of interest. Now that we have understood the principle of the DIF construction, it is easy to use this principle in order to obtain a parameterized probability distribution with a tractable PDF by selecting (i) a base distribution $\mathcal{Q}$ with tractable density, (ii) classification functions w_k (i.e. one that computes a categorical vector of probability), and (iii) K changes of variables T_k.

By construction, the DIF extends mixture models, but they can also be seen as an extension of NF in which the deterministic mapping is replaced by a discrete stochastic one. Mixtures of NF have already been proposed in the literature and have been mildly successfully applied in some contexts. In our work, we do not focus on this aspect but rather investigate whether or not we can leverage the flexibility induced by the mixture weights $\pi_k(x)$ (or equivalently $w_k(z)$). To that end, we consider a DIF with a standard normal base distribution $\mathcal{Q} = \mathcal{N}(\mu = \mathbf{0}_d, \Sigma = \mathbf{I}_d)$, with $T_k^{-1}(z) = \mu_k + \Sigma_k^{1/2}z$, and with functions $w_k(z)$ defined via an NN classifier function. The corresponding DIF therefore has a PDF which reads:

$$p_{\text{DIF}}(x) = \sum_{k=1}^K w_k(\Sigma_k^{-1/2}x - \mu_k)\mathcal{N}(x; \mu_k, \Sigma_k); \tag{4.41}$$

which is effectively an extension of a GMM where the constant mixture weights are replaced by a function of X.

4.2 Reparameterization gradient

Most models presented in the previous section are indeed generative models in the sense that they describe a sampling procedure, which can usually be straightforward. As such, they can be used to estimate expectations with MC as:

$$\mathbb{E}_{X_\theta \sim \mathcal{P}_\theta}\big[g_\theta(X_\theta)\big] \overset{MC}{\approx} = \frac{1}{M}\sum_{i=1}^{M} g_\theta(x_\theta^{(i)}), \text{ where } x_\theta^{(1)}, ..., x_\theta^{(M)} \overset{iid}{\sim} \mathcal{P}_\theta. \tag{4.42}$$

However, an auxiliary problem that occurs in many instances of statistical modeling, is that of computing (or estimating) the gradient of the quantity $\mathbb{E}_{X_\theta \sim \mathcal{P}_\theta}\big[g_\theta(X_\theta)\big]$. In this section, we specifically denote $X_\theta \in \mathbb{R}^d$ with subscript θ to stress that it indeed depends on the parameter θ as a $\mathcal{P}_\theta$ distributed RV. As we have explained several times thus far, such expectations rarely admit closed-form expression, which calls for Monte-Carlo estimation using iid samples drawn from $\mathcal{P}_\theta$. However, the samples indirectly depend on parameter $\theta \in \mathbb{R}^n$, the parameters of the probability distribution with respect to which we compute the expectation. Therefore, computing an estimate of:

$$\nabla_\theta \Big(\mathbb{E}_{X_\theta \sim \mathcal{P}_\theta}\big[g_\theta(X_\theta)\big] \Big)|_{\theta=\theta_t}; \tag{4.43}$$

is not straightforward (36)(74)(84) (see (77) for a survey). Eliciting a parametric family $\mathcal{P}_\theta$ such that this gradient is easy to estimate (see, for instance, (55)(61)); or finding an appropriate estimation method for a given model ((38) for mixtures, (65) for Gaussian distributions) is of crucial importance. We consider g_θ a measurable function which we want to compute the expectation of under $\mathcal{P}_\theta$, and is also denoted with the subscript θ since it may as well depend on θ. It is the case, for example, in a VI context when we want to minimize $D_{\mathrm{KL}}(\mathcal{P}_\theta || \mathcal{P})$ via Gradient Descent (GD), in which case the function reads $g_\theta(.) = \log\Big(\frac{p_\theta(.)}{p(.)}\Big)$. We suppose in this section that the function g_θ is differentiable with respect to its argument and with respect to parameters θ as well, so that the gradient is well defined. We first come to the realization that this problem reduces to computing a gradient of the form:

$$\nabla_\theta \Big(\mathbb{E}_{X_\theta \sim \mathcal{P}_\theta}\big[f(X_\theta)\big] \Big)|_{\theta=\theta_t}, \tag{4.44}$$

where f does not depend on θ. Indeed, under mild conditions which enable permuting integral and gradient, (4.43) becomes:

$$\begin{aligned} &\nabla_\theta \Big(\mathbb{E}_{X_\theta \sim \mathcal{P}_\theta}\big[g_\theta(X_\theta)\big] \Big)|_{\theta=\theta_t} = \\ &\mathbb{E}_{Z_{\theta_t} \sim q_{\theta_t}}\big[\nabla_\theta g_\theta(Z_{\theta_t})|_{\theta=\theta_t}\big] + \nabla_\theta \Big(\mathbb{E}_{X_\theta \sim \mathcal{P}_\theta}\big[g_{\theta_t}(X_\theta)\big] \Big)|_{\theta=\theta_t}. \end{aligned} \tag{4.45}$$

In this expression, the first term can be computed easily so long as g_θ can be differentiated with respect to θ (which we suppose is the case). Hence, all the burden of computing (4.43) is computing the second term which, if we set $f = g_{\theta_t}$, is nothing but (4.44) and the goal of this section is to explain how one can estimate such a gradient.

A first approach is to use the principle of importance sampling to rewrite the expectation computed with respect to $\mathcal{P}_\theta$ into an expectation computed with respect to a probability distribution $\mathcal{Q}$ which does not depend on θ:

$$\mathbb{E}_{X_\theta \sim \mathcal{P}_\theta}\big[f(X_\theta)\big] = \mathbb{E}_{Z\sim\mathcal{Q}}\bigg[f(Z)\frac{p_\theta(Z)}{q(Z)}\bigg]. \tag{4.46}$$

Then, computing the gradient of (4.44) reduces to computing the gradient of the previous expression (4.46), which, since the probability distribution with respect to which the expectation is computed no longer depends on θ, can easily be done with permutation of the gradient and expectation:

$$\nabla_\theta \mathbb{E}_{X_\theta \sim \mathcal{P}_\theta}\big[f(X_\theta)\big]|_{\theta=\theta_t} = \mathbb{E}_{Z\sim\mathcal{Q}}\bigg[\frac{f(Z)}{q(Z)}\nabla_\theta p_\theta(Z)|_{\theta=\theta_t}\bigg]. \tag{4.47}$$

It is straightforward to build a MC estimate of this gradient using iid. samples $z_1, ..., z_M$ from the importance distribution $\mathcal{Q}$ as:

$$\nabla_\theta \mathbb{E}_{X_\theta \sim \mathcal{P}_\theta}\big[f(X_\theta)\big]|_{\theta=\theta_t} \approx \frac{1}{M}\sum_{i=1}^{M}\frac{f(z_i)}{q(z_i)}\nabla_\theta p_\theta(z_i)|_{\theta=\theta_t}. \tag{4.48}$$

In importance sampling, the variance of the estimate is determined not only by the discrepancy between the distribution of interest and the importance distribution but also by the variations of f. In the case of using the importance sampling principle as a mean to estimate the gradient, it is still unclear how to appropriately select the importance distribution (i) to improve the corresponding estimate for fixed t and (ii) in a sequential scheme as t increases, as is the case in a gradient-based optimization scheme. This question might be the topic of future work. However, a most notable particular case is the *log-trick*, also named the *reinforce gradient* method since it was popularized in a reinforcement learning policy adjustment setting (86), corresponds to using $\mathcal{P}_{\theta_t}$ as an importance distribution. With the fact that $\frac{\nabla_\theta p_\theta(.)|_{\theta=\theta_t}}{p_{\theta_t}(.)} = \nabla_\theta \log(p_\theta(.))|_{\theta=\theta_t}$, the expression for the gradient then becomes:

$$\nabla_\theta\Big(\mathbb{E}_{X_\theta\sim\mathcal{P}_\theta}\big[f(X_\theta)\big]\Big)|_{\theta=\theta_t} = \mathbb{E}_{X_{\theta_t}\sim\mathcal{P}_{\theta_t}}\Big[f(X_{\theta_t})\nabla_\theta \log\big(p_\theta(X_{\theta_t})\big)|_{\theta=\theta_t}\Big];$$

which, once again, can easily be estimated with Monte-Carlo using iid. samples drawn from the $\mathcal{P}_{\theta_t}$. This approach is therefore very simple to implement in a sequential optimization scheme, as it simply requires drawing samples from the current model and can always be used as long as the PDF $p_\theta(.)$ can be evaluated and differentiated. It can even be applied to discrete probability distributions as opposed to the reparameterization trick, which is not necessarily compatible with all probability distributions, as we will see in the next section. However, several works, such as (66) seem to conclude that

the reinforce log-trick produces high variance estimates, especially when compared to the reparameterization trick (88), but examples in (33) show that this is not always the case.

In the previous section, we built an estimate of the gradient of the expectation of interest (4.44) by first rewriting it into an expectation computed with respect to a distribution that does not depend on parameters θ and we were able to do so via the importance sampling principle. This can also be achieved by using a change of variable, which is the principle behind the *reparameterization trick* (50)(53): consider a bivariate function $T(.;.)$ such that RV $T(X_\theta;\theta)$ is distributed from a distribution which does not depend on θ. T is therefore often called a standardization function. We denote ϵ that RV and $\mathcal{Q}$ the probability distribution according to which it is distributed. For instance, if $T(.;\theta)$ is the CDF of $\mathcal{P}_\theta$, then ϵ is a uniform RV $\epsilon \sim \mathcal{U}\big[0,1\big]^d$ which indeed does not depend on θ. If T is invertible and differentiable with respect to its first argument so that it defines a valid change of variables and is differentiable with respect to θ (which then forces T^{-1} to be as well); then, we can rewrite (4.44) using the change of variables technique for integrals with $X_\theta = T^{-1}(\epsilon;\theta)$:

$$\mathbb{E}_{X_\theta\sim\mathcal{P}_\theta}\big[f(X_\theta)\big] = \mathbb{E}_{\epsilon\sim\mathcal{Q}}\big[f(T^{-1}(\epsilon;\theta))\big]. \tag{4.49}$$

Then, since the expectation is computed with respect to a distribution which no longer depends on θ, one can compute its gradient easily with permutation of gradient and expectation:

$$\nabla_\theta\Big(\mathbb{E}_{X_\theta\sim\mathcal{P}_\theta}\big[f(X_\theta)\big]\Big)|_{\theta=\theta_t} = \mathbb{E}_{\epsilon\sim\mathcal{Q}}\Big[\nabla_x f(x)|_{x=T^{-1}(\epsilon;\theta_t)}\nabla_\theta T^{-1}(\epsilon;\theta)|_{\theta=\theta_t}\Big]. \tag{4.50}$$

Note that (4.50) holds regardless of the function T. By hypothesis, we can easily compute the gradient of f with respect to z, therefore, computing (4.43) reduces to computing the gradient of a sample X_θ with respect to θ, and an estimate of the gradient can be computed with MC as:

$$\frac{1}{M}\sum_{i=1}^{M}\nabla_x f(x)|_{x=T^{-1}(\epsilon_i;\theta_t)}\nabla_\theta T^{-1}(\epsilon_i;\theta)|_{\theta=\theta_t} \text{ where } \epsilon_1,...,\epsilon_M \sim \mathcal{Q} \tag{4.51}$$

Therefore, the standard approach for applying a reparameterization trick consists of a two-step procedure. First, one needs to exhibit $T(.;.)$ (with adequate differentiability and invertibility) such that $\epsilon = T(X_\theta;\theta)$ does not depend on θ. Second, one needs to invert $T(.;\theta)$ and compute its gradient. However, in some cases, a standardization function T exists and can be computed, but can not be inverted in an analytical form. This is the case, for instance, with gamma distributions, where numerical methods exist to compute the CDF, but its inverse can not be computed. In this context, (32) proposes to rewrite the gradient as:

$$\nabla_\theta T^{-1}(\epsilon;\theta)|_{\theta=\theta_t} = -\Big(\nabla_z T(z;\theta_t)|_{z=T^{-1}(\epsilon;\theta_t)}\Big)^{-1}\nabla_\theta T(T^{-1}(\epsilon;\theta_t);\theta)|_{\theta=\theta_t}; \tag{4.52}$$

which then enables to rewrite 4.50 as:

$$\begin{aligned}&\nabla_\theta\Big(\mathbb{E}_{X_\theta\sim\mathcal{P}_\theta}\big[f(X_\theta)\big]\Big)|_{\theta=\theta_t} = \\ &-\mathbb{E}_{X_{\theta_t}\sim\mathcal{P}_{\theta_t}}\Big[\nabla_x f(x)|_{x=X_{\theta_t}}\Big(\nabla_x T(x;\theta_t)|_{x=X_{\theta_t}}\Big)^{-1}\nabla_\theta T(X_{\theta_t};\theta)|_{\theta=\theta_t}\Big].\end{aligned} \tag{4.53}$$

This expression can therefore easily be estimated by MC with using samples from $\mathcal{P}_{\theta_t}$ and does not require explicit computation of T^{-1}. More recently, other alternative gradients estimation methods have been proposed: namely (45)(76)(3) which propose generalization of the reparameterization trick. More specifically, other works (see (28) for reference) have successfully proposed reparamerization gradient estimation of rejection-sampling scheme (62), MCMC sampling (83) and slice sampling (89).

Finally, in this context of gradient estimation, the culmination of this chapter with the DIF construction does provide an easy reparameterization gradient, even though its sampling procedure is non-differentiable as it involves a categorical (discrete) RV. In the bulk of this chapter, we present the corresponding gradient estimate in the specific context of gradient-based parameter adjustment for VI as a specific Rao-Blackwellization (RB) procedure.

4.3 Discretely Indexed Flows

In this section we propose DIF as a new tool for solving variational estimation problems. Roughly speaking, DIF are built as an extension of NFs, in which the deterministic transport becomes stochastic, and more precisely discretely indexed. Due to the discrete nature of the underlying additional latent variable, DIF inherit the good computational behavior of NF: they benefit from both a tractable density as well as a straightforward sampling scheme, and can thus be used for the dual problems of VI and of VDE. On the other hand, DIF can also be understood as an extension of mixture density models, in which the constant mixture weights are replaced by flexible functions. As a consequence, DIF are better suited for capturing distributions with discontinuities, sharp edges and fine details, which is a main advantage of this construction. Finally we propose a methodology for constructiong DIF in practice, and see that DIF can be sequentially cascaded, and cascaded with NF.

4.3.1 Introduction

Many scientific tasks take interest in decision making with respect to some random process. In this context, evaluating the PDF and/or obtaining random samples from the process can help the decision making by computing statistical quantities of interest. For example computing confidence intervals may help to conclude on the existence or absence of some underlying effect. Historical methods include posterior inference with MCMC (79) (19) or Approximate Bayesian Computation in the likelihood-free setting (5).

The task of probabilistic modeling provides with a concurrent approach: by using an approximating distribution (sometimes referred to as a *surrogate*) with either or both a tractable density and an explicit sampling mechanism, we can estimate relevant statistics. This includes the non-parametric approach of Kernel Density Estimation (69). Variational probabilistic modeling consists in building a surrogate probability distribution by solving an optimization problem among some parametric family of distributions (41) (8) (24). Recent advances in automatic differentiation (4) (70) (1) and optimiza-

tion (49) have paved the way to using NNs functions in probabilistic modeling (50) (37) (80). Note however that concurrent approaches can perform density estimation (39) with leveraging NN functions which approximate a density ratio but without explicitly constructing a probability distribution.

NFs (54) (67) are a versatile tool for probabilistic modeling as they allow for both generation of random samples with an explicit sampling mechanism, and density estimation with exact PDF evaluation. Therefore NF are at the crossroad between VI (73) (52), VDE (26) (68) and Generative modeling (51). These three problems are especially relevant in the field of machine learning which explain the popularity of NF amongst machine learning practitionners. Moreover, part of their attractiveness results from the fact that NF define deterministic invertible transformations which can effortlessly be layered to produce deep and flexible families of surrogates, making them competitive on a performance standpoint.

In this work, we build DIF as an extension of NF, and we therefore provide another method in order to build surrogate probability distributions. DIF no longer rely on a deterministic mapping but rather leverage a stochastic transformation, all the while remaining in the same sweet spot as NF: they allow for both exact PDF evaluation and straightforward sampling. On the other hand, DIF can also be seen as an extension of mixture density models, in which the constant mixture weights become flexible functions. As a result, DIF enable to capture distributions with finer details than regular mixture models.

The rest of this section is organised as follows. In section 4.3.2 we present the two dual problems of VI, on the one hand, and VDE, on the other hand. Both are probabilistic modeling problems, in which we build a surrogate $\mathcal{P}_\theta$ of the true probability distribution $\mathcal{P}$; in the first case, we use the PDF p associated to $\mathcal{P}$, and in the second case observed samples from $\mathcal{P}$. In section 4.3.3, we recall the principles of NF, explain how they can be used for VI and VDE, and revisit them as latent variables models.

In section 4.3.5 we extend NF to DIF; roughly speaking, the deterministic transport is replaced by a (discrete) stochastic one, therefore DIF are LVMs too, but the original latent space is augmented by an additional discrete variable. From a computational point of view, DIF retain the good behaviour of NF; indeed, the discrete nature of the additional variable enables for explicit density evaluation as well as a closed form formula for the reverse transition kernel between the latent and observed spaces. We next see that similarly to NF, DIF can be used efficiently either for the VI or for the VDE problems; as far as VI is concerned, our work builds upon the previous Transport Monte Carlo (TMC) approach (29), but we argue in favor of a more coherent optimization objective than that used in the TMC approach.

Finally in section 4.3.8 we propose a methodology for constructing DIF in practice. Namely we propose a convenient parameterization of the DIF stochastic transport. Under this parameterization, DIF can be considered as an extension of a GMM, the benefits of which are illustrated via simulations on complex two-dimensional distributions as well as compared to elaborate NFs in the context of learning a distribution over images [1]. We finally see that DIF can be combined together (or with NF as well), and that they

[1]We provide all reproducible code and experiments in the Github repository at github.com/ElouanARGOUARCH/Discretely-Indexed-Flows.

can be used for conditional density estimation.

4.3.2 Two dual probabilistic modeling problems

In this section we propose a parallel discussion of the VI and VDE problems, which are the two modeling problems addressed by the DIF methodology.

Variational Inference

Suppose that we dispose of $p(x)$, in a possibly unnormalized form, but we do not have a simple procedure for sampling from the distribution $\mathcal{P}$. This is usually the case when considering a posterior distribution, the PDF of which is proportional to the product of the *prior* and of the *likelihood*, but the normalizing constant (the *evidence*) is unavailable. VI aims at providing samples that are approximately distributed according to $\mathcal{P}$, by considering a variational distribution defined as:

$$\mathcal{P}_\theta^\star = \arg\min_{\mathcal{P}_\theta} \mathrm{D}^{(\mathrm{VI})}\big(\mathcal{P}_\theta, \mathcal{P}\big),$$

where $\mathrm{D}^{(\mathrm{VI})}$ is some discrepancy measure and $\mathcal{P}_\theta$ belongs to some family of distributions which is straightforward to sample from. Since $\mathcal{P}_\theta^\star$ is close to $\mathcal{P}$, samples from $\mathcal{P}_\theta^\star$ are approximately distributed according to P.

Note moreover that if the PDF $p_\theta^\star$ is available, one can use $\mathcal{P}_\theta^\star$ as an importance distribution for targeting $\mathcal{P}$. Furthermore, one can use Rubin's sampling importance-resampling mechanism (22) (34) (78) (14, §9.2) to produce asymptotically independent and identically distributed (iid) samples from $\mathcal{P}$.

We consider $\mathrm{D}^{(\mathrm{VI})}$ to be the Kullback-Leibler Divergence (56) (D_{KL}), be it either the forward one $D_{\mathrm{KL}}(\mathcal{P}||\mathcal{P}_\theta)$ or the reverse one $D_{\mathrm{KL}}(\mathcal{P}_\theta||P)$. We also consider a parametric family $\big\{\mathcal{P}_\theta|\theta \in \Theta\big\}$. However for arbitrary $\mathcal{P}$, neither $D_{\mathrm{KL}}(\mathcal{P}||\mathcal{P}_\theta)$ nor $D_{\mathrm{KL}}(\mathcal{P}_\theta||P)$ admits a closed form expression, which calls for a Monte Carlo (MC) approximation. Since we can only sample from $\mathcal{P}_\theta$, the discrepancy measure $\mathrm{D}^{(\mathrm{VI})}$ must be the reverse D_{KL}, and an MC approximation can be computed as:

$$D_{\mathrm{KL}}(\mathcal{P}_\theta||P) = \mathbb{E}_{\mathcal{P}_\theta}\left(\log\left(\frac{p_\theta(X)}{p(X)}\right)\right) \approx \frac{1}{M}\sum_{\substack{i=1\\ x_i\sim\mathcal{P}_\theta}}^{M} \log\left(\frac{p_\theta(x_i)}{p(x_i)}\right). \tag{4.54}$$

Minimizing this MC estimate with respect to model parameters θ via GD requires that p is differentiable (which is assumed throughout this work), and also that p_θ is chosen to be differentiable with respect to θ. However computing gradients can still be challenging because the samples $x_i \sim \mathcal{P}_\theta$ indeed depend on model parameter θ. One way to compute the gradients is to use a *reparameterization trick*, that is, to use an invertible differentiable standardization function $S(.;\theta)$ such that RV $S(X;\theta) = \epsilon$, does not depend on θ. Then re-writing x_i as $x_i = S^{-1}\big(\epsilon_i;\theta\big)$ enables computing the gradients of (4.54) with respect to θ.

Variational Density Estimation

Suppose that we dispose of samples $x_1, ..., x_M \sim \mathcal{P}$ but we cannot evaluate the PDF $p(x)$. This occurs for example when we have only recorded observations from an otherwise unknown real-world stochastic process. Among other techniques, we can perform VDE to obtain an estimation of $p(x)$ by considering a variational distribution defined as:

$$\mathcal{P}_\theta^\star = \arg\min_{\mathcal{P}_\theta} \mathrm{D}^{(\mathrm{VDE})}\big(\mathcal{P}_\theta, \mathcal{P}\big),$$

where $\mathrm{D}^{(\mathrm{VDE})}$ is some discrepancy measure and $\mathcal{P}_\theta$ belongs to some family of distributions with tractable PDF. Since $\mathcal{P}_\theta^\star$ is close to $\mathcal{P}$, the PDF $p_\theta^\star$ is an estimate of the unknown density function p.

Note moreover that if $\mathcal{P}_\theta^\star$ is easy to sample from, then samples from $\mathcal{P}_\theta^\star$ are approximately distributed according to $\mathcal{P}$.

Once again, we will consider $\mathrm{D}^{(\mathrm{VDE})}$ to be a D_{KL} and the parametric family $\big\{\mathcal{P}_\theta | \theta \in \Theta\big\}$. Minimizing an MC approximation of the reverse D_{KL}, as in (4.54), is not possible here. Indeed in this case, the PDF p is evaluated at samples points $x_i \sim \mathcal{P}_\theta$, which depend on θ; hence we need to account for the terms $p(x_i)$ in the optimization, which is not possible since function $p(.)$ is unknown. This calls for the use of the *forward* D_{KL}, and an MC approximation using the samples from $\mathcal{P}$ can be computed as :

$$D_{\mathrm{KL}}(\mathcal{P}||\mathcal{P}_\theta) = \mathbb{E}_{X\sim\mathcal{P}}\left(\log\left(\frac{p(X)}{p_\theta(X)}\right)\right) \approx \frac{1}{M}\sum_{\substack{i=1\\ x_i\sim\mathcal{P}}}^{M} \log\left(\frac{p(x_i)}{p_\theta(x_i)}\right).$$

Though this MC estimate of the D_{KL} cannot be computed since p is not available, note that x_i and $p(x_i)$ do not depend on θ, and can thus be ignored in the minimization process. As a consequence minimizing this MC approximation of the D_{KL} reduces to maximizing the log-likelihood of the data under model $\mathcal{P}_\theta$. Finally, minimizing the MC approximation of $D_{\mathrm{KL}}(\mathcal{P}||\mathcal{P}_\theta)$ can be conducted via GD, which only requires that p_θ is differentiable (here, unlike in section 4.3.2 the samples do not depend on θ, so the gradients can be computed directly).

4.3.3 Normalizing Flows

In this section we propose a brief presentation of NF, which have been first introduced in (73) (see also (54) (67) for thorough reviews of the topic). We explain how to use NF for the two problems of VI and VDE.

Change of variables, sampling mechanism and density evaluation

The underlying idea of NF is that of a bijective change of variables. Let U and V be two RVs related via:

$$V = f^{-1}(U) \tag{4.55}$$

for some C1-diffeomorphism f, that is, an invertible mapping such that both f and its inverse f^{-1} are differentiable and with continuous derivatives. Let q_U and q_V be

respectively the PDF of U and V. As is well known, (4.55) induces:

$$q_V(x) = q_U\big(f(x)\big)\big|\det \mathrm{J}_f(x)\big|, \tag{4.56}$$

where J_f is the Jacobian matrix. This change of variables formula for densities in fact defines the PDF q_V via a functional transform $\mathcal{F}$ of q_U and of the mapping f:

$$q_V = \mathcal{F}\big(q_U; f\big) \tag{4.57}$$

These formulas are potentially useful for sampling (4.55) and for density evaluation (4.56). However, at this point it is interesting to observe that they do not involve the same assumptions on q_U and f:

1. If f^{-1} is available without tears, and if q_U is easy to sample from, then (4.55) can be used as a straightforward sampling mechanism: if $u \sim q_U$ then $v = f^{-1}(u) \sim q_V$, in other words we first sample u from q_U and then map u to v via f^{-1};

2. On the other hand, evaluating q_V via (4.56) requires that PDF q_U can be evaluated at any point and that we can compute the Jacobian determinant easily.

Application to the variational problems

Let us now see how to apply (4.55) and (4.56) to the variational problems identified in section 4.3.2. Given a $\mathcal{Q}$ distributed RV Z (usually chosen as a fixed standard Gaussian distribution $\mathcal{N}\big(0, \mathrm{I}_d\big)$ - as discussed in section 4.3.3 - where d is the dimension of the problem), both problems consist in designing a change of variables T such that the distribution of $\widetilde{X} = T^{-1}(Z)$, which we denote as $\mathcal{P}_\theta$, is close to target $\mathcal{P}$ (in the sense of the appropriate D_{KL}). In the general case and for both problems of VDE and VI, the optimization problem will not have a solution for arbitrary T. Therefore, we consider $T \in \big\{T_\theta | \theta \in \Theta\big\}$, with the condition that T is differentiable with respect to model parameters θ, in order to solve the optimization using GD (see (25) (26) (68) (52) (51) (59) (30) for examples of such parametrization).

The fact that mapping T is a C1-diffeomorphism has interesting consequences. First, mapping T indeed provides two couples of rv: $\Big[\widetilde{X} = T_\theta^{-1}(Z), Z\Big]$ (see the second row of figure 4.2), but also $\Big[X, \widetilde{Z} = T_\theta(X)\Big]$ (see first row), in which $X \sim \mathcal{P}$ and $\mathcal{Q}_\theta$ denotes the distribution of $\widetilde{Z}$. Then, if $\mathcal{Q}$ and $\mathcal{P}$ respectively admit PDF q and p, the PDF p_θ

$$\begin{array}{ccc} x - obs. & & z - lat. \\ P & \xrightarrow{T_\theta(X)=\widetilde{Z}} & \mathcal{Q}_\theta \\ \mathcal{P}_\theta & \xleftarrow{\widetilde{X}=T_\theta^{-1}(Z)} & Q \end{array}$$

Figure 4.2: Forward & Backward mappings between observed and latent spaces

and q_θ associated with $\mathcal{P}_\theta$ and $\mathcal{Q}_\theta$ are defined via the same functional transform (4.57):

$$p_\theta = \mathcal{F}\big(q; T_\theta\big) \tag{4.58}$$
$$q_\theta = \mathcal{F}\big(p; T_\theta^{-1}\big) \tag{4.59}$$

Moreover, by applying the simple change of variables $z = T_\theta(x)$, we have the following two equalities:

$$D_{\mathrm{KL}}\big(\mathcal{P}_\theta||\mathcal{P}\big) = D_{\mathrm{KL}}\big(Q||\mathcal{Q}_\theta\big), \tag{4.60}$$
$$D_{\mathrm{KL}}\big(\mathcal{P}||\mathcal{P}_\theta\big) = D_{\mathrm{KL}}\big(\mathcal{Q}_\theta||Q\big). \tag{4.61}$$

These two equalities explain that, since observed and latent distributions are related via a deterministic invertible mapping, minimizing a forward (resp. reverse) D_{KL} in the latent space (that of $\widetilde{Z}$ and Z) mechanically minimizes a reverse (resp. forward) D_{KL} in the observed space (that of X and $\widetilde{X}$), and vice-versa. These remarks will be useful in later sections.

Why *Normalizing*? Why *Flow*?

- With an argument similar to the inverse CDF technique for sampling (67, §2.2), for any distribution $\mathcal{P}$ with compact support, we can (at least theoretically) construct a transport T between $\mathcal{P}$ and a standard Normal distribution (whence the term *Normalizing* Flows). For that reason, it is routinely assumed to set $\mathcal{Q}$ as a standard parameter-free Normal distribution (which is assumed from now on); and using an NF for modeling P reduces to approximating a combination of two CDF.

- In order to ensure sufficient flexibility in T, we leverage the property that C1-diffeomorphisms are closed under composition. If we define for example $T = T_1 \circ T_2$, where T_1, T_2 are C1-diffeomorphisms, then T is also a C1-diffeomorphism and its Jacobian determinant can be computed using the chain rule formula:

  $$|\det \mathrm{J}_{T_\theta}(x)| = |\det \mathrm{J}_{T_1}(x)| \times |\det \mathrm{J}_{T_2}\big(T_1(x)\big)|,$$

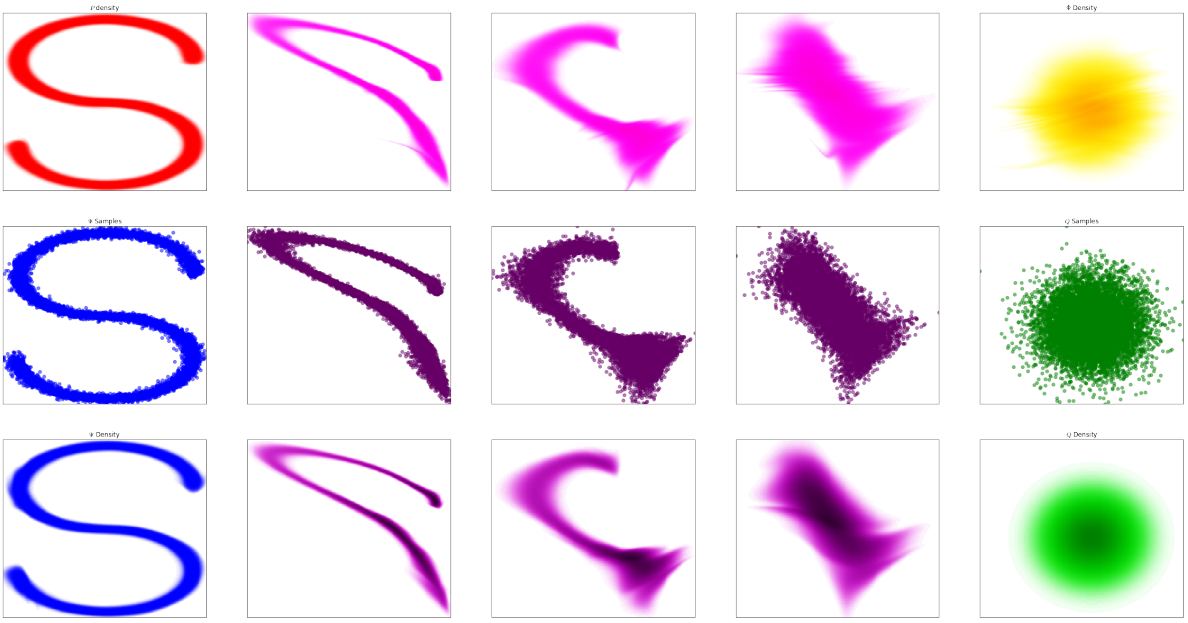
  which implies $\mathcal{F}\big(.; T\big) = \mathcal{F}\Big(\mathcal{F}\big(.; T_2\big); T_1\Big)$. Hence, it is easy to construct T as a composition of simple transformations $\{T_c\}_{c=1,\dots,C}$. The distribution $\mathcal{Q}$ sequentially gets morphed into $\mathcal{P}_\theta$ by a *Flow* of transformations, whence the term Normalizing *Flows*.

VI with NF

Let us consider the VI setting described in section 4.3.2. Minimizing an MC approximation of the reverse D_{KL} leads to the following optimization problem:

$$\min_{\theta\in\Theta} \sum_{\substack{i=1 \\ x_i\sim\mathcal{P}_\theta}}^{M} \log\left(\frac{p_\theta(x_i)}{p(x_i)}\right). \tag{4.62}$$

Figure 4.3: Example of VI using a multi-step NF



In order for this objective to be differentiated with respect to θ we may try to apply a reparametrization trick. Equation (4.60) hints at a solution for such a reparameterization: since $x_i = T^{-1}(z_i)$, where $z_i \sim \mathcal{Q}$ does not depend on θ (because we have assumed that $\mathcal{Q}$ is parameter-free) and T is differentiable with respect to model parameters θ. Hence by construction, NF provides with a straightforward differentiable reparametrization trick $x = T^{-1}(z)$. By applying this change of variable, (4.62) becomes:

$$\begin{aligned}\min_{\theta\in\Theta} \sum_{\substack{i=1\\ z_i\sim\mathcal{Q}}}^{M} \log\left(\frac{p_\theta\big(T^{-1}(z_i)\big)}{p\big(T^{-1}(z_i)\big)}\right) &= \min_{\theta\in\Theta} \sum_{\substack{i=1\\ z_i\sim\mathcal{Q}}}^{M} \log\left(\frac{q(z_i)}{q_\theta(z_i)}\right) \\ = \max_{\theta\in\Theta} \sum_{\substack{i=1\\ z_i\sim\mathcal{Q}}}^{M} \log\big(q_\theta(z_i)\big) &\overset{(4.59)}{=} \max_{\theta\in\Theta} \sum_{\substack{i=1\\ z_i\sim\mathcal{Q}}}^{M} \log\left(p\big(T^{-1}(z_i)\big)\left|\det \mathrm{J}_{T_\theta^{-1}}(z_i)\right|\right). \end{aligned} \tag{4.63}$$

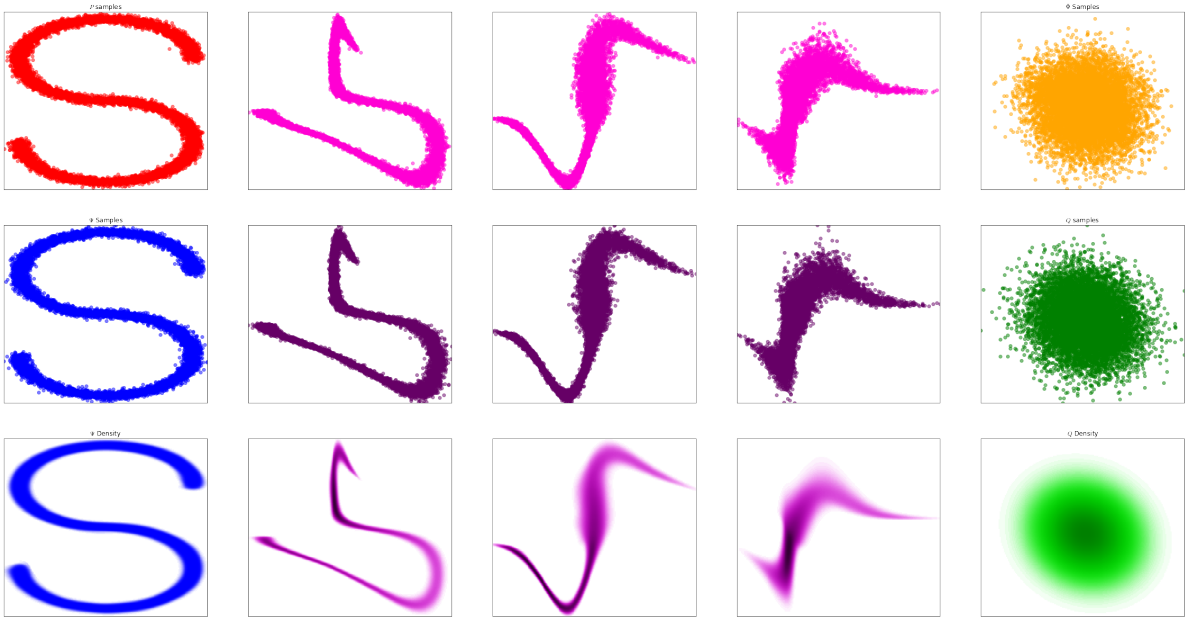
The resulting optimization problem can be solved using GA as this expression is differentiable with respect to θ, and $\mathcal{Q}$ was purposely chosen to be easy to sample from. Let $\theta^\star$ maximize (4.63); it remains to sample the corresponding model $\mathcal{P}_\theta^*$ to produce samples that are approximately distributed according to $\mathcal{P}$.

Figure 4.3 presents an example of an NF used for VI on a 2-dimensional S-Curve problem. The left most column shows the observed-x space while the right most column corresponds to the latent-z space. The model is defined as a composition of 4 Real-NVP coupling layers (26), the middle columns present the intermediate distributions between each transformation. We can therefore visualise how a standard Gaussian distribution $\mathcal{Q}$ (green) is sequentially morphed into the model distribution $\mathcal{P}_\theta$ (blue) that resembles the target $\mathcal{P}$ (red). As expected, $\mathcal{Q}_\theta$ (yellow) resembles $\mathcal{Q}$. The first row shows a color-mapping of the target density p function getting morphed via T. The last two rows

are both representations of $\mathcal{P}_\theta$ but the first shows drawn samples while the later is a color-mapping of the density function p_θ which is a result of q getting morphed via T^{-1}.

DE with NF

Figure 4.4: Example of Density Estimation and Sampling using a multi-step NF



Consider now the VDE setting described in section 4.3.2, the MLE problem (which we recall is equivalent to minimizing an MC approximation of the forward D_{KL}) reads:

$$\max_{\theta\in\Theta} \sum_{\substack{i=1\\ x_i\sim\mathcal{P}}}^{M} \log\big(p_\theta(x_i)\big) \overset{(4.58)}{=} \max_{\theta\in\Theta} \sum_{\substack{i=1\\ x_i\sim\mathcal{P}}}^{M} \Big(q\big(T_\theta(x)\big)\big|\det \mathrm{J}_{T_\theta}(x)\big|\Big) \tag{4.64}$$

For this optimization objective to be differentiable, it is only necessary that the density function q is also differentiable, which is the case since $\mathcal{Q}$ is a standard Normal distribution. Then, the maximization can be solved using GA. Let $\theta^\star$ maximize equation (4.64); then model PDF $p^*_\theta(x)$ is an approximation of the target PDF p and hence solves the VDE problem. Note moreover that we can produce new samples that are approximately distributed from $\mathcal{P}$ by sampling the corresponding model $\mathcal{P}^*_\theta$.

Figure 4.4 presents an example of an NF used for VDE on the same target distribution as in figure 4.3. The flow model is also defined as a composition of 4 Real NVP layers. The only difference with Figure 4.3 is that the target distribution is available via its samples and therefore the first row shows the samples of $\mathcal{P}$ being transformed via T, the interpretation of this figure is otherwise the same.

Topological limitations

Observe however that, by essence, NF are not well suited to approximate multimodal distributions with disjoints supports: since T^{-1} continuously reshapes the Normal dis-

tribution $\mathcal{Q}$ into $\mathcal{P}_\theta$, the model will struggle to efficiently split the mass into several modes. We illustrate this in figure 4.5 where we try using a multi-step NF to approach a distribution with two disjoint moon elements. In this context, we see that there remains an artefact connection between the two elements of mass; hence the resulting NF distribution is not one with disjoint supports. This topological limitation is one drawback of NF models which DIF circumvent, see section 4.3.10 below.

Figure 4.5: Topological limitation of an NF; it struggles to approach distribution with disjoint support

4.3.4 From Normalizing Flows to Discretely Indexed Flows

As we now see, NF can be considered as LVMs, which suggests the extension to DIF which will be addressed in section 4.3.5.

Flows as LVMs

NF target a distribution $\mathcal{P}$ by constructing a distribution $\mathcal{P}_\theta$ associated with a RV $\widetilde{X} = T_\theta^{-1}(Z)$. The RV Z is a proxy latent variable distributed according to $\mathcal{Q}$, and we adjust T so that $\mathcal{P}_\theta$ is as close as possible of $\mathcal{P}$. $\mathcal{P}_\theta$ is therefore the marginal distribution of interest, out of a couple of RVs $(\widetilde{X}, Z)$ with joint density $\overleftarrow{\pi_\theta}(x|z)q(z)$. It can thus be considered as an LVM: we are given a prior q (the distribution of the latent variable Z), and we move from z to x via the conditional distribution $x \sim \overleftarrow{\Pi_\theta}(.|z)$. Of course, since Z and $\widetilde{X} = T_\theta^{-1}(Z)$ are related via a *deterministic* mapping, the associated conditional density function $\overleftarrow{\pi_\theta}(x|z)$ reads:

$$\overleftarrow{\pi_\theta}(x|z) = \delta_{T_\theta^{-1}(z)}(x). \tag{4.65}$$

Beyond Normalizing Flows

One way of increasing expressiveness is to consider an LVM in which the deterministic mapping (4.65) used in NF is replaced by a stochastic transport described by a transition kernel $\overleftarrow{\Pi_\theta}$, be it a PDF or a probability mass function. In either case, due to the latent variable structure, sampling from $\mathcal{P}_\theta$ is (almost) as easy as in the deterministic case: we start off by sampling the latent distribution $z \sim \mathcal{Q}$, and next we sample from the conditional distribution $x \sim \overleftarrow{\Pi_\theta}(.|z)$. Therefore, as long as the prior and likelihood are both easy to sample from, the model can be sampled from effortlessly.

The density associated with $\mathcal{P}_\theta$ is given by:

$$p_\theta(x) = \mathbb{E}_{Z\sim\mathcal{Q}}\left[\overleftarrow{\pi_\theta}(x|Z)\right]$$

This density is not necessarily tractable as the expectation does not always admit a closed form expression, at least if the conditional distribution is continuous. If however $\overleftarrow{\pi_\theta}(x|z)$ is discrete and has finite support, then the integral becomes a tractable sum.

As a consequence, note that for continuous LVMs, we might not be able to use explicit evaluation of the density function p_θ in order to optimize the model. For example, if p_θ is untractable, we cannot perform direct MLE of model parameters and we have to rely on more sophisticated optimization procedures. For instance (50) presents the VAE, which is a continuous LVM that leverages a variational EM optimization scheme.

However, for the task of VDE, it is desirable that we use a model $\mathcal{P}_\theta$ s.t. PDF p_θ exists and can be computed exactly. Therefore, we will explore DIF which is a class of LVMs that builds upon the principle of invertible mappings used in NF and includes stochasticity in a discrete form to ensure tractable density.

4.3.5 The DIF construction

In this section we introduce DIF as one possible stochastic extension of NF. More precisely, we build DIF as an LVM where the deterministic mapping between z and x is replaced by a discrete stochastic distribution. We then apply DIF for both the VI and VDE problems.

DIF as a discrete LVM

We define a DIF model via some prior distribution $\mathcal{Q}$ and the likelihood:

$$\overleftarrow{\pi_\theta}(x|z) = \sum_{k=1}^{K} w_k(z)\delta_{T_k^{-1}(z)}(x), \tag{4.66}$$

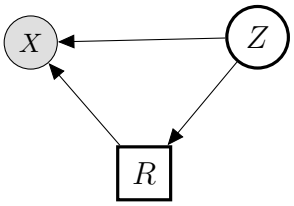
where, though we drop the subscript notation $\{w_k, T_k\}_{k=1,\dots,K}$, are specified functions parameterized by θ (see section 4.3.8 to understand how such functions can be parameterized). With words: for a given value of z, z is transformed into $x = T_k^{-1}(z)$ via

mapping T_k with probability $w_k(z)$. The function $w_k(z)$ therefore represents the conditional probability $\Pr\big(x = T_k^{-1}(z)|z\big)$ and must sum to 1: $\sum_{k=1}^{K} w_k(z) = 1, \forall z \in \mathbb{R}^d$.

DIF in fact is an auxiliary LVM where we leverage a categorical latent variable U to create a stochastic transport instead of a deterministic one. The latent RV U takes discrete values 1 to K which indicate what mapping is applied to z. The resulting RV $\widetilde{X}$ can be written as :

$$\widetilde{X} = T_R^{-1}(Z) \text{ where } Z \sim \mathcal{Q} \text{ and } R \sim \text{Categorical}\big(w_1(Z), ..., w_K(Z)\big). \tag{4.67}$$

Hence, sampling from this model remains straightforward as the stochastic transport of prior samples $z \sim \mathcal{Q}$ can be conducted with sampling $\overleftarrow{\Pi}_\theta(.|Z = z)$ using (4.67). DIF is therefore a viable parametric model candidate for VI (see section 4.3.6 below).



With this choice for $\overleftarrow{\Pi}_\theta$ and with the additional constraint that each T_k for $k = 1, ..., K$ is a C1-diffeomorphism, one can show (see appendix A.4) that the marginal PDF p_θ reads

$$p_\theta(x) = \sum_{k=1}^{K} w_k\big(T_k(x)\big) q\big(T_k(x)\big) \big|\det \mathrm{J}_{T_k}(x)\big|. \tag{4.68}$$

Once again, this PDF is therefore defined as a functional transform of the prior PDF q and of functions $\{w_k, T_k\}_{k=1,...,K}$, which we denote similarly as

$$p_\theta = \mathcal{F}\Big(q; \overleftarrow{\Pi}_\theta\Big). \tag{4.69}$$

Since (4.68) can be computed in closed-form, DIF can be used to tackle VDE (see section 4.3.7 below).

At this point, let us observe that DIF can be seen as an extension of two different classes of models:

- For $K = 1$, the stochastic transform becomes deterministic, and indeed the DIF reduces to an NF;
- A DIF with $K > 1$ components but with constant functions w_k is nothing but a mixture model, since in this case the categorical latent variable U does not depend on z. This point of view will prove of particular interest in section 4.3.8.

It is also important to note that extending a deterministic transport to a discrete stochastic transport has two main advantages. The first one is that the prior mass can be sent to different regions which enables to effectively split the prior mass. Second, this enables to use a function

Back and forth between observed and latent space

Recall that for NF the transport was deterministic and invertible, so we were able to go back and forth between observed and latent spaces by applying either T or T^{-1} (see section 4.3.3).

In the case of DIF, for a given value z, x is one of the values $\{T_1^{-1}(z), ..., T_K^{-1}(z)\}$ with associated probabilities $\{w_1(z), ..., w_K(z)\}$; and similarly, for a given x, z must be one of the values $\{T_1(x), ..., T_K(x)\}$ with associated probabilities $\{v_1(x), ..., v_K(x)\}$. Indeed one can show (see appendix A.4) that the forward transport $\overrightarrow{\Pi_\theta}$ reads:

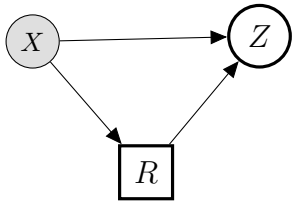
$$\begin{aligned} \overrightarrow{\pi_\theta}\big(z|x\big) &= \sum_{k=1}^{K} v_k(x)\delta_{T_k(x)}(z), \\ v_k(x) &= \frac{w_k\big(T_k(x)\big)q\big(T_k(x)\big)\big|\det \mathrm{J}_{T_k}(x)\big|}{\sum_{j=1}^{K} w_j\big(T_j(x)\big)q\big(T_j(x)\big)\Big|\det \mathrm{J}_{T_j}(x)\Big|}. \end{aligned} \tag{4.70}$$

We now define the RV $\widetilde{Z}$ (and denote $\mathcal{Q}_\theta$ its probability distribution):

$$\widetilde{Z} = T_R(X) \text{ where } X \sim \mathcal{P} \text{ and } R \sim \text{Categorical}\big(v_1(X), ..., v_K(X)\big). \tag{4.71}$$

So using DIF is almost as convenient as using NF: even though the transportation $\overleftarrow{\Pi_\theta}$ from z to x is stochastic, $\overrightarrow{\Pi_\theta}$ remains of the same (discrete) nature. The interest of this result is twofold:

- If we dispose of $x \sim \mathcal{P}$, we can easily obtain a sample from $\mathcal{Q}_\theta$ by applying (4.71).



Therefore, since we can sample easily from both the likelihood (backward transport $\overleftarrow{\Pi_\theta}$) and the posterior (forward transport $\overrightarrow{\Pi_\theta}$), we can go back and forth between observed and latent spaces just like in NF. The following diagram summarizes the discussion, and should be compared to figure 4.2.

$$\begin{array}{ccc} x - obs. & & z - lat. \\ \mathcal{P} & \xrightarrow{\overrightarrow{\Pi_\theta}(.|X)\sim\widetilde{Z}} & \mathcal{Q}_\theta \\ \mathcal{P}_\theta & \xleftarrow{\widetilde{X}\sim\overleftarrow{\Pi_\theta}(.|Z)} & Q \end{array}$$

Figure 4.6: Forward & Backward transitions between obs. and lat. spaces

- One can show easily that
$$q_\theta = \mathcal{F}\Big(p; \overrightarrow{\Pi_\theta}\Big). \tag{4.72}$$
Similarly to NF, both model densities p_θ and q_θ can thus be written as the result of the same functional transform (compare (4.69) to (4.72), (4.69) to (4.58) and (4.72) to (4.59)). Moreover, if we dispose of the PDF p associated with $\mathcal{P}$, we can compute q_θ explicitly.

Note that $\mathcal{F}(q_\theta; \overleftarrow{\Pi_\theta}) \neq p$, which we can relate to the fact that the backward transportation in the $(X, \widetilde{Z})$ joint distribution is not $\overleftarrow{\Pi_\theta}$. Instead, for purpose of later arguments, let us denote $\overleftarrow{\Pi_\theta}'$ the backward transport such that $\mathcal{P} = \mathcal{F}(q_\theta; \overleftarrow{\Pi_\theta}')$ and hence:

$$p(x)\overrightarrow{\pi_\theta}(z|x) = \overleftarrow{\pi_\theta}'(x|z)q_\theta(z).$$

Related work and comparison to the TMC approach

In this section we briefly recall some alternate extensions of NF which have been introduced previously. In section 4.3.5 we particularly focus on the TMC approach, which is closely related to our work; this section will be of particular interest in section 4.3.6, where we will further extend the comparison between the two approaches under the scope of the VI problem.

There have been several prior works which attempted at constructing extensions of NF by using non-deterministic transformations.

Continuously Indexed Flows (CIF) consider a hierarchical LVM with a continuous indexing latent variable. CIF have been applied to both VI in (16) and generative modeling settings (20). However, due to the continuous nature of the augmenting rv, CIF do not admit a tractable density.

Augmented Normalizing Flows (ANF) (42) augment the observation with a continuous RV and use a deterministic NF in order to learn the joint density. ANF produce an augmented likelihood which does not allow for exact density evaluation. Both CIF and ANF are classes of models which include the VAE (50) and, due to their intractable density, cannot be trained via direct likelihood maximization. Instead, just like VAE, the training consists in maximizing the likelihood via a variational EM scheme. SurVAE (63) aims at providing a unified framework for building complex generative models with the use of surjective and stochastic transformations (of which CIF, ANF and DIF are instances). Perhaps more closely related to DIF, (27) considers a piecewise invertible flow-type transformation which also corresponds to using a discrete indexing variable, but where the induced partitioning is hard. This approach allows for a tractable density and does not require summing over the discrete indexing variable since only one of the component is non-negative for any observation.

In particular, our work can be connected to the previous TMC approach (29). As we shall see in this section, though DIF and TMC use similar stochastic constructions, we will argue in favor of DIF which can be applied to both problems of VI and VDE, while TMC is only suited for the VI setting. Moreover, specifically in a VI setting, the TMC approach considers a particular optimization objective. In section 4.3.6 we will continue the comparison between DIF and TMC in order to discuss the motivation of

this optimization objective, and finally in section 4.3.6 we will propose a more coherent optimization objective which can be applied to both DIF and TMC.

The TMC approach is closely related to the methodology proposed by DIF in the sense that it considers a stochastic transport of the same nature as DIF. The definition of the model is however done in a different order as compared to DIF. Indeed, in the TMC approach, the starting point is the target PDF $p(x)$ to which we apply a forward stochastic transport of the form $\overrightarrow{\pi_\theta}(z|x) = \sum_{k=1}^K v_k(x)\delta_{T_k(x)}(z)$. We therefore obtain a joint distribution $(X, \widetilde{Z})$ with marginal $q_\theta(z)$ given by (4.72). From this joint distribution, we can computed the associated backward transport:

$$\begin{aligned} \overleftarrow{\pi_\theta}(x|z) &= \sum_{k=1}^K w_k(z)\delta_{T_k^{-1}(z)}(x), \\ w_k(z) &= \frac{v_k\big(T_k^{-1}(z)\big)p\big(T_k^{-1}(z)\big)\Big|\det \mathrm{J}_{T_k^{-1}}(z)\Big|}{\sum_{j=1}^K v_j\Big(T_j^{-1}(z)\Big)p\Big(T_j^{-1}(z)\Big)\Big|\det \mathrm{J}_{T_j^{-1}}(z)\Big|}. \end{aligned} \tag{4.73}$$

Finally, the model distribution in the observed space is defined as $\mathcal{P}_\theta$ which corresponds to $\mathcal{Q}$ transported via the the backward transport $\overleftarrow{\Pi_\theta}$. Note again that the forward transportation in $(\widetilde{X}, Z)$ is not $\overrightarrow{\Pi_\theta}$ and we instead denote $\overrightarrow{\Pi_\theta}'$.

$$\begin{array}{ccc} x-obs. & & z-lat. \\ \mathcal{P} & \underset{X\sim\overleftarrow{\Pi_\theta}(.|\widetilde{Z})}{\overset{\overrightarrow{\Pi_\theta}(.|X)\sim\widetilde{Z}}{\leftrightarrows}} & \mathcal{Q}_\theta \\ \mathcal{P}_\theta & \underset{\overrightarrow{\Pi_\theta}'(.|\widetilde{X})\sim Z}{\overset{\widetilde{X}\sim\overleftarrow{\Pi_\theta}(.|Z)}{\leftrightarrows}} & \mathcal{Q} \end{array} \qquad \begin{array}{ccc} x-obs. & & z-lat. \\ \mathcal{P} & \underset{X\sim\overleftarrow{\Pi_\theta}'(.|\widetilde{Z})}{\overset{\overrightarrow{\Pi_\theta}(.|X)\sim\widetilde{Z}}{\leftrightarrows}} & \mathcal{Q}_\theta \\ \mathcal{P}_\theta & \underset{\overrightarrow{\Pi_\theta}(.|\widetilde{X})\sim Z}{\overset{\widetilde{X}\sim\overleftarrow{\Pi_\theta}(.|Z)}{\leftrightarrows}} & \mathcal{Q} \end{array}$$

Figure 4.7: TMC approach (left) compared to DIF (right) - parameterized transports are indicated in red

To summarize, the TMC approach considers a forward transport $\overrightarrow{\Pi_\theta}$ between $\mathcal{P}$ and $\mathcal{Q}_\theta$, then computes its backward transport $\overleftarrow{\Pi_\theta}$ which is finally applied to $\mathcal{Q}$ in order to obtain the model $\mathcal{P}_\theta$. DIF and TMC are in fact defined in reverse order compared to one another since in the DIF approach, we consider a backward transport $\overleftarrow{\Pi_\theta}$ between $\mathcal{Q}$ and $\mathcal{P}_\theta$, and we consequently deduce the forward transportation $\overrightarrow{\Pi_\theta}$ which can then be applied to $\mathcal{P}$ in order to obtain $\mathcal{Q}_\theta$. As a consequence, though the expressions for $p_\theta(x), q_\theta(z), \overleftarrow{\pi_\theta}(x|z)$ and $\overrightarrow{\pi_\theta}(z|x)$ look similar and are obtained via similar computations, in the TMC approach we set v_k and compute w_k while in the DIF approach we set w_k and compute v_k.

In section 4.3.6 we will discuss the pros and cons of each approach under the scope of the VI problem, but at this point let us already notice that TMC cannot be applied

to a density estimation setting. Indeed, with this choice of parameterization, w_k given in (4.73) is computed from v_k but also depends on the density p which is not available in the density estimation setting. On the other hand, in the DIF approach, functions w_k are parameterized directly and do not depend on PDF p. Consequently, we can compute directly w_k and $p_\theta(x)$, which enables sampling and density evaluation in both settings of VI and VDE. Finally, we can already argue in favor of DIF since it provides with a more versatile tool for tackling both variational problems.

4.3.6 Application of DIF to VI

So far, we have presented the general principles of DIF and explained that it consists in a natural extension of NF. In particular, even though the transformation is now stochastic, DIF defines a model PDF p_θ which remains computable, and also provides the ability to go back and forth between observed and latent spaces.

In this section we discuss the use of a DIF for tackling a VI problem and we therefore consider the setting described in 4.3.2. As we have mentionned in the previous section 4.3.5, DIF and TMC are closely related. However, TMC considers a particular optimization objective function. In section 4.3.6 and 4.3.6 we further compare TMC to DIF in order to discuss the relevance of this optimization objective, and finally in section 4.3.6 we argue in favor of a more motivated optimization objective.

Computational aspects

With the same notations introduced before, TMC builds a model $\mathcal{P}_\theta$ for $\mathcal{P}$ by solving the optimization problem:

$$\max_{\theta\in\Theta} \sum_{\substack{i=1 \\ z_i\sim\mathcal{Q}}}^{M} \log\big(q_\theta(z_i)\big), \tag{4.74}$$

which corresponds to minimizing an MC approximation of $D_{\mathrm{KL}}\big(Q||\mathcal{Q}_\theta\big)$.

However, the optimization problem (4.74) seems strange at first sight, because the standard approach for VI would indeed prescribe minimizing a discrepancy between $\mathcal{P}_\theta$ and $\mathcal{P}$ (see (4.54)). But minimizing an MC approximation of $D_{\mathrm{KL}}\big(\mathcal{P}_\theta||\mathcal{P}\big)$ would yield the optimization objective:

$$\min_{\theta\in\Theta} \sum_{\substack{i=1 \\ x_i\sim\mathcal{P}_\theta}}^{M} \log\left(\frac{p_\theta(x_i)}{p(x_i)}\right). \tag{4.75}$$

Since the samples $x_i \sim \mathcal{P}_\theta$ depend on model parameters θ, we should apply a reparameterization trick, that is, write x_i as $x_i = T^{-1}(\epsilon_i;\theta)$ in which $\epsilon_i \sim \epsilon$ and rv ϵ does not depend on θ. In the case of NF (see section 4.3.3), the deterministic mapping automatically induced a differentiable reparameterization trick $x_i = T^{-1}(z_i)$. Here by contrast, since sampling from $\mathcal{P}_\theta$ involves sampling from an auxiliary categorical latent variable, finding an invertible change of variable which is differentiable with respect to θ is likely not to be possible. Therefore, the minimization problem (4.75) cannot be conducted via GD.

By contrast, the objective function in (4.74) is easy to maximize: sampling from $\mathcal{Q}$ is straightforward by design, and this objective is differentiable with respect to θ and can therefore be maximized via GA. This computational argument argues in favor of the optimization objective in the TMC approach, but on the other hand, one can wonder whether minimizing a discrepancy in the latent space induces similar counterparts in the observed space, see section 4.3.6 below.

Variational aspects

Unfortunately, by contrast with NF, the two equalities between D_{KL} (4.60) and (4.61) no longer hold when we work with a DIF or a TMC model. Therefore it is not so obvious that minimizing a D_{KL} in the latent space, that is between $\mathcal{Q}$ and $\mathcal{Q}_\theta$, produces a good approximation of $\mathcal{P}$ with $\mathcal{P}_\theta$. Nonetheless, we can justify to some extent the use of this optimization objective in TMC. Indeed, we notice that the forward D_{KL} in the latent space is an upper bound of the reverse D_{KL} in the observed space:

$$D_{\mathrm{KL}}\big(\mathcal{Q}||\mathcal{Q}_\theta\big) = D_{\mathrm{KL}}\big(\mathcal{P}_\theta||\mathcal{P}\big) + \mathbb{E}_{X\sim\mathcal{P}_\theta}\left[D_{\mathrm{KL}}\Big(\overleftarrow{\Pi}_\theta(.|X)||\overleftarrow{\Pi}_\theta{}'(.|X)\Big)\right] \geq D_{\mathrm{KL}}\big(\mathcal{P}_\theta||\mathcal{P}\big).$$

Note moreover that we have similarly $D_{\mathrm{KL}}\big(\mathcal{Q}_\theta||\mathcal{Q}\big) \geq D_{\mathrm{KL}}\big(\mathcal{P}||\mathcal{P}_\theta\big)$. Therefore, the TMC approach minimizes (an MC approximation of) an upper bound of the usual optimization objective (4.54) defined in the VI setting.

Since D_{KL} are positive, it follows that if a D_{KL} between $\mathcal{Q}_\theta$ and $\mathcal{Q}$ (forward or reverse) reaches zero via optimization, both forward and reverse D_{KL} between $\mathcal{P}_\theta$ and $\mathcal{P}$ reach zero. However, forcing a D_{KL} in the latent space to zero means that the prior PDF q belongs to $\big\{q_\theta|\theta\in\Theta\big\}$, or equivalently that there exists $\overleftarrow{\Pi}_\theta$ and q such that $\mathcal{F}\Big(q;\overleftarrow{\Pi}_\theta\Big) = p$ (see (4.69)), which is unlikely to be the case for arbitrary distributions $\mathcal{P}$. Moreover, standard optimization techniques such as GD only guarantee convergence to a local extremum of the objective function. So in practice we have to deal with positive D_{KL} in the latent space as we may only reach a local minimum of a positive function. There is furthermore no evidence that a local minimum of D_{KL} in the latent space is also a local minimum of D_{KL} in the observed space. Finally we cannot conclude with certainty that a decent approximation of $\mathcal{Q}$ with $\mathcal{Q}_\theta$ (in the D_{KL} sense) produces a good model $\mathcal{P}_\theta$ and we would preferably want to obtain a minimum of D_{KL} in the observed space.

In the case of the DIF, since the model is defined the other way round as compared to TMC, the majorization obtained for TMC becomes a minorization:

$$D_{\mathrm{KL}}\big(\mathcal{P}_\theta||\mathcal{P}\big) = D_{\mathrm{KL}}\big(\mathcal{Q}||\mathcal{Q}_\theta\big) + \mathbb{E}_{Z\sim\mathcal{Q}}\left[D_{\mathrm{KL}}\Big(\overrightarrow{\Pi}_\theta(.|Z)||\overrightarrow{\Pi}_\theta{}'(.|Z)\Big)\right] \geq D_{\mathrm{KL}}\big(\mathcal{Q}||\mathcal{Q}_\theta\big).$$

Therefore, in the case of DIF, minimizing a latent D_{KL} would only minimize a lower bound of the observed D_{KL} which we should minimize in the VI setting. This consideration would argue against DIF if we were not able to minimize directly the D_{KL} in the observed space. Fortunately, as we now see, it is indeed possible in both the DIF and TMC cases to minimize directly an MC approximation of the $D_{\mathrm{KL}}(\mathcal{P}_\theta||\mathcal{P})$.

Rao-Blackwellizing the D_{KL} estimate

From sections 4.3.6 and 4.3.6, we see that it would be desirable to minimize a discrepancy measure in the observed space. Could we write an estimate of $D_{\mathrm{KL}}(\mathcal{P}_\theta||\mathcal{P})$ for which we are able to compute the gradients with respect to model parameters θ ? Such a case would be ideal, since we could perform GD while ensuring that the model converges toward a local minimum of the discrepancy measure in the observed space.

It happens that the following estimator :

$$D_{\mathrm{KL}}\big(\mathcal{P}_\theta||\mathcal{P}\big) \approx \frac{1}{M}\sum_{\substack{i=1\\ z_i\sim\mathcal{Q}}}^{M}\sum_{k=1}^{K} w_k(z_i)\log\left(\frac{p_\theta\big(T_k^{-1}(z_i)\big)}{p\big(T_k^{-1}(z_i)\big)}\right) \tag{4.76}$$

is one possible solution since, by contrast with (4.75), this D_{KL} estimate is differentiable with respect to model parameters. As we now see, this estimate indeed comes as the result of an RB procedure (15) (34) (75). Let J be the rv

$$J = \log\left(\frac{p_\theta\Big(\widetilde{X}\Big)}{p\Big(\widetilde{X}\Big)}\right) \overset{(4.67)}{=} \log\left(\frac{p_\theta\big(T_R^{-1}(Z)\big)}{p\big(T_R^{-1}(Z)\big)}\right),$$

where $Z\sim\mathcal{Q}$ and $R\sim\mathrm{Categorical}\big(w_1(Z),...,w_K(Z)\big)$. It is clear that

$$D_{\mathrm{KL}}(\mathcal{P}_\theta||\mathcal{P}) = \mathbb{E}(J),$$

where the expectation is taken with respect to the joint distribution of $\big(Z,R\big)$. On the one hand, sampling $\big\{(z_i,r_i)\big\}_{i=1,...,M}$ from this joint distribution yields the crude MC estimate in (4.75). On the other hand, RB is based on the observation that $\mathbb{E}(J) = \mathbb{E}\Big(\mathbb{E}\big(J|Z\big)\Big)$ (7). Since we can compute the inner expectation:

$$\mathbb{E}\big(J|Z\big) = \sum_{k=1}^{K} w_k(Z)\log\left(\frac{p_\theta\big(T_k^{-1}(Z)\big)}{p\big(T_k^{-1}(Z)\big)}\right),$$

only the outer one calls for an MC approximation, so we only need to sample $z_i\sim\mathcal{Q}$. This leads to the estimator $D_{\mathrm{KL}}(\mathcal{P}_\theta||\mathcal{P}) \approx \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\big(J|Z=z_i\big)$, which is nothing but (4.76).

The interest of using this RB estimate is twofold. First, as is well known $\mathbb{V}\mathrm{ar}\Big(\mathbb{E}\big(J|Z\big)\Big) = \mathbb{V}\mathrm{ar}(J) - \mathbb{E}\Big(\mathbb{V}\mathrm{ar}\big(J|Z\big)\Big)$, so (4.76) has lower variance than the estimator in (4.75). Next, and more importantly in our context, we are no longer reliant on a reparameterization of the Categorical RV U, so the estimate is now differentiable. Indeed, before resorting to an MC approximation, we have computed whatever could be computed, namely $\mathbb{E}\big(J|Z\big)$ (where the expectation is taken with respect to U). Therefore the estimate does not involve sampling from U since this RV has been explicitly marginalized out.

4.3.7 Application of DIF for VDE

As we have seen already, DIF are designed such that they can also be used for VDE, since the backward transport and density p_θ do not depend on p (remember that p is not available in a density estimation setting, see section 4.3.2). In this section we now explain precisely how DIF can be used for the VDE problem.

The MLE approach

As we now see, the issues that were identified when using DIF for VI no longer occur when tackling the problem of VDE with a DIF model. To see this, suppose that we dispose of samples $x_i \sim \mathcal{P}$. Since p_θ can be written in closed form, the MLE problem reads

$$\max_{\theta\in\theta} \sum_{\substack{i=1\\ x_i\sim\mathcal{P}}}^{M} \log\big(p_\theta(x_i)\big) \overset{(4.68)}{=} \max_{\theta\in\theta} \sum_{\substack{i=1\\ x_i\sim\mathcal{P}}}^{M} \log\left(\sum_{k=1}^{K} w_k\big(T_k(x_i)\big) q\big(T_k(x_i)\big) \big|\det \mathrm{J}_{T_k}(x_i)\big|\right). \quad (4.77)$$

By contrast with (4.75), x_i are sampled from $\mathcal{P}$, and not from $\mathcal{P}_\theta$, so do not depend on θ. As a result, we see from the right hand side of (4.77) that this objective is differentiable with respect to θ.

A Generalized EM (GEM) procedure

Let us turn to computational optimization aspects. Of course, the objective function in (4.77) can be directly optimized with GD in an Automatic Differentiation framework like Pytorch or Tensorflow. However, let us observe that this function involves the logarithm of a sum, which leads to entangled gradients. As a consequence, gradients computation could be slowed down.

In this section we propose an alternative optimization procedure, based on a MM (43)(82) approach, the principle of which is as follows. Instead of optimizing directly a function $f(\theta)$, we sequentially build a series of surrogate functions $\big\{g_{\theta_1}(\theta), g_{\theta_2}(\theta), \cdots\big\}$ which locally minorate f and for which $g_{\theta_t}(\theta_t) = f(\theta_t)$. From this series of functions, one can sequentially deduce a series of parameters $\big\{\theta_1, \theta_2, ...\big\}$ such that $g_{\theta_t}(\theta_{t+1}) \geq g_{\theta_t}(\theta_t)$. So

$$f(\theta_{t+1}) \geq g_{\theta_t}(\theta_{t+1}) \geq g_{\theta_t}(\theta_t) = f(\theta_t),$$

which finally ensures that $f(\theta_t)$ converges to a local maximum of $f(\theta)$. Moreover, if both $f(\theta)$ are $g_{\theta_t}(\theta)$ are differentiable functions, then the gradients evaluated at $\theta = \theta_t$ must be equal. To see this, let us consider the function $\theta \mapsto f(\theta) - g_{\theta_t}(\theta)$; in the vicinity of θ_t, this function is non-negative, differentiable and is zero for $\theta = \theta_t$. Hence θ_t is a local minimum and its gradient is zero:

$$\nabla_\theta f(\theta)|_{\theta=\theta_t} = \nabla_\theta g_{\theta_t}(\theta)|_{\theta=\theta_t}. \quad (4.78)$$

We now apply this technique to the above optimization problem. We obtain the surrogate function $g_{\theta_t}(\theta)$ for the likelihood function in (4.77) (we omit variables x_i in

the left hand side of (4.79) since the samples are fixed); details are given in appendix A.5:

$$\begin{aligned} g_{\theta_t}(\theta) &= \sum_{i=1}^{M}\sum_{k=1}^{K} v_k^{(\theta_t)}(x_i)\log\left(\frac{h_k^{(\theta)}(x_i)}{v_k^{(\theta_t)}(x_i)}\right), \\ h_k^{(\theta)}(x_i) &= w_k^{(\theta)}\Big(T_k^{(\theta)}(x_i)\Big)q\Big(T_k^{(\theta)}(x_i)\Big)\Big|\det \mathrm{J}_{T_k^{(\theta)}}(x_i)\Big|. \end{aligned} \tag{4.79}$$

If we perform GA with respect to θ , we obtain a GEM scheme (87), which ensures that the model converges toward a local maximum in (4.77). Finally observe that our surrogate function no longer involves a log-sum but rather a sum-log, which detangles the computation of its gradients.

4.3.8 Using DIF in practice

In this section we first explain how one can parameterize the functions T_k and w_k to produce an efficient DIF model. We then propose an overall view of the DIF mechanism, as a stochastic transport transforming $\mathcal{Q}$ into the resulting distribution $\mathcal{P}_\theta$. We then revisit DIF as an extension of mixture density models but where the constant weights are replaced by an arbitrary function of z; we illustrate this effect on complex distributions, and see that DIF enable to capture sharp edges and finer details as compared to a standard GMM. We also see that complex DIF can be constructed as the succession of simpler building blocs: in the same spirit as NF, we propose an approach for cascading DIF layers. Finally we discuss the use of DIF in the specific setting of conditional density estimation, and see that one could easily turn a DIF into a conditional density model.

Example of DIF parameterization

As we have explained before, using DIF requires solving an optimization problem (be it for the VI or VDE problems), in which the multidimensional parameter θ gathers those of the probability functions w_k, as well as of the invertible mappings T_k. Even though this optimization is performed with respect to the parameters altogether, w_k and T_k still play a different role and must be specified accordingly.

The functions $w_k(.)$ are straightforward to parameterize, since the only constraints are that these functions are differentiable, non negative, and sum to 1 for any given input vector z. This can be achieved by defining $w_1(z), ..., w_K(z)$ as the output of a K-label classifier architecture with input z by computing the unnormalized weights $\widetilde{w_1}(z), ..., \widetilde{w_K}(z)$ and applying a softmax normalization. This ensures that the weights sum to one and form a valid vector of categorical probabilities. More precisely, we can for instance consider the architecture of a multi-Layer percepetron, which is an NN function with L hidden layers, each layer $l = 1, ..., L$ having n_l hidden units:

$$\begin{aligned} &h_1 = \sigma(W_0 z + b_0); \\ &h_{l+1} = \sigma(W_l h_l + b_l) \text{ for } l = 1, ..., L-1; \\ &\big[\widetilde{w_1}(z), ..., \widetilde{w_K}(z)\big]^T = W_L h_L + b_L; \\ &\big[w_1(z), ..., w_K(z)\big]^T = \mathrm{Softmax}\Big(\big[\widetilde{w_1}(z), ..., \widetilde{w_K}(z)\big]^T\Big), \end{aligned}$$

where $W_l \in \mathbb{R}^{n_{l+1} \times n_l}$ and $b_l \in \mathbb{R}^{n_{l+1}}$ for $l = 0, ..., L$ (with $n_0 = d$ and $n_{L+1} = K$) are the weights parameters, and where $\sigma(.)$ is some chosen element-wise activation function (for example the sigmoid function).

When selecting the parametric functions T_k, we actually dispose of a wide range of possibilities. Depending on the problem, the only constraint is that the functions must be changes of variables, and that T_k^{-1} (for VI) or the Jacobian determinant (for VDE) can be computed easily. We may consider simple location-scale mappings like in GMM, or we may borrow from the NF literature such as in (25) (26) (68) (52) (51) (30) (59). In section 4.3.10 we propose a construction which reduces the burden of parameterizing the mappings T_k. Moreover, if we consider weights w_k defined via a flexible parametric function (as in section 4.3.8), we do not require flexible invertible mappings T_k to produce a flexible DIF. Therefore we only consider here a simple location-scale transformation

$$T_k^{-1}(z) = m_k + s_k \odot z, \tag{4.80}$$

where $m_k \in \mathbb{R}^d$ (a translation vector, whence the term *location*) and $s_k \in \mathbb{R}_+^{*\,d}$ (a *scale* vector) are the parameters be optimized, and $\odot$ is the element wise vector product. Since s_k is strictly positive, T_k is invertible and we can easily obtain the inverse mapping as well as the Jacobian determinant with:

$$T_k(x) = s_k^{-1} \odot (x - m_k) \text{ and } \big|\det \mathrm{J}_{T_k}(x)\big| = \prod_{j=1}^{d} \Big(s_k^{(j)}\Big)^{-1},$$

where s_k^{-1} is the vector of element-wise inverses of s_k. Note moreover that, in order to easily ensure that s_k remains positive throughout a gradient based optimization, we parameterize and optimize $\log(s_k)$.

Generalizing standard GMM

Under this parameterization, and due to the discrete latent structure of DIF, we in fact obtain a model which can be compared to a GMM. Indeed, if $\mathcal{Q}$ is a multivariate Normal distribution, then we have:

$$q\big(T_k(x)\big)\big|\det \mathrm{J}_{T_k}(x)\big| = \mathcal{N}\big(x; m_k, D_k\big),$$

with D_k the diagonal matrix with values are the squared elements of s_k. Hence, the corresponding model density (4.68) reads:

$$p_\theta(x) = \sum_{k=1}^{K} w_k\big(s_k^{-1} \odot (x - m_k)\big)\mathcal{N}\big(x; m_k, D_k\big), \tag{4.81}$$

and can be interpreted as the density associated with a mixture of Gaussian distributions with diagonal covariance matrices, in which the constant mixture weights are replaced by functions of x which read $w_k\big(s_k^{-1} \odot (x - m_k)\big)$. Note that, while in a standard GMM the constant mixture weights sum to 1, it is not necessarily the case with DIF. Indeed, though we have that $\sum_{k=1}^{K} w_k(z) = 1$ for any $z \in \mathbb{R}^d$, this does not

imply that $\sum_{k=1}^{K} w_k\big(s_k^{-1} \odot (x - m_k)\big)$ is 1. This sum is equal to 1 in two particular cases. First, when all the mappings are the same, i.e. when $m_1 = m_2 = ... = m_K$ and $s_1 = s_2 = ... = s_K$, in which case the model reduces to a unique Gaussian distribution. Second, when the functions w_k are constant, in which case we retrieve a standard GMM.

A main advantage of such a parameterization is that the parameters m_k and s_k can be interpreted as the means and standard variation vectors of the different Gaussian components while the functions w_k are responsible for weighting and modulating each component. Here the Gaussian components are isotropic since, along each dimension j, the latent variable z has its element $z^{(j)}$ scaled by $s_k^{(j)}$. We could also have considered a full covariance matrix transformation in order to obtain an analogous model of a full covariance matrix GMM by replacing (4.80) by a transformation of the type $T_k^{-1}(z) = m_k + S_k z$ where S_k is a square root of an invertible covariance matrix. However, this leads to the burden of parameterizing S_k in a way which ensures that, throughout a gradient based optimization, $S_k^T S_k$ remains an invertible matrix.

In the next section 4.3.8, we illustrate the DIF mechanism, and we explain that making the mixture weights dependent on the variable x increases the flexibility of the underlying model as compared to the standard (constant weight) GMM.

An overall view: the DIF de- and re-constructs $\mathcal{Q}$ into $\mathcal{P}_\theta$

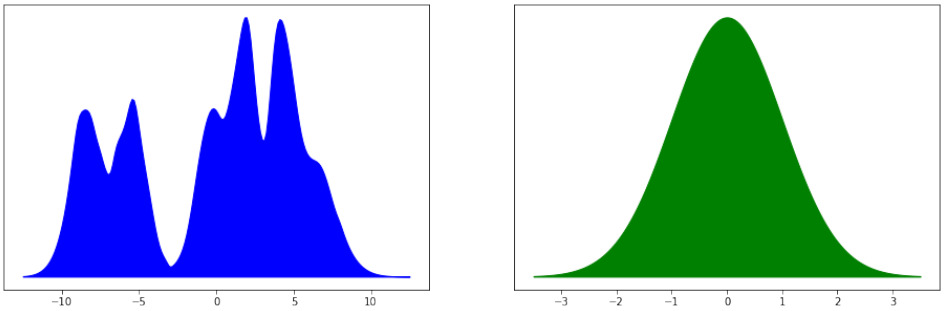
We now illustrate via a one-dimensional example how a DIF transforms the prior distribution $\mathcal{Q}$ into a complex probability distribution $\mathcal{P}_\theta$ with density given by (4.68) (see figure 4.8a), and in particular explain the roles of weights $\{w_k\}_{k=1,...,K}$ and of mappings $\{T_k\}_{k=1,...,K}$. The discussion in this section is of course independent of the problem tackled (VI or VDE) and of the associated optimization.

First, the left hand side of figure 4.8b displays the weight functions $w_k(z)$. Since they are positive and sum to 1 (for any z), we have $q(z) = \sum_{k=1}^{K} w_k(z) q(z)$; so functions $w_k(z)$ induce a soft partitioning of the latent space, and indeed split the prior mass into several parts, see the right hand side of figure 4.8b (or equivalently the first row of figure 4.8c). These figures indeed provide a way of visualizing the joint distribution (Z, R): the values of z can be read on the x-axis, and the values of $U = 1, ..., K$ are the different colors in the right hand side of figure 4.8b (or the different sub-figures in figure 4.8c).

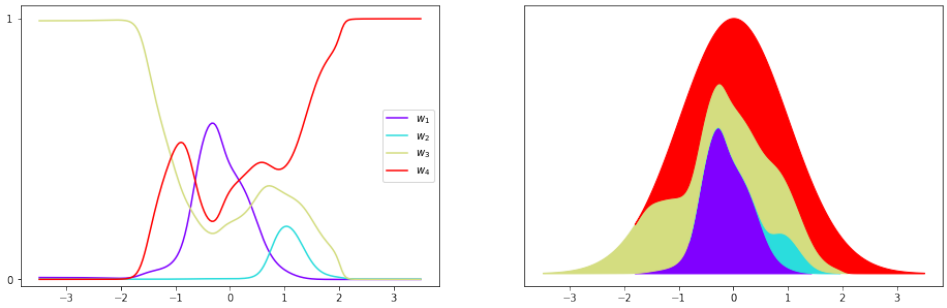
Next, given $U = k$, a prior sample z is transported via mapping T_k with $x = T_k^{-1}(z)$. So on the whole, the C1-diffeomorphisms T_k^{-1} send the elements of mass in possibly different regions of the observed space, and continuously reshape them (see second row of figure 4.8c).

Finally all these parts are recombined into the final probability distribution $\mathcal{P}_\theta$, see figure 4.8d.

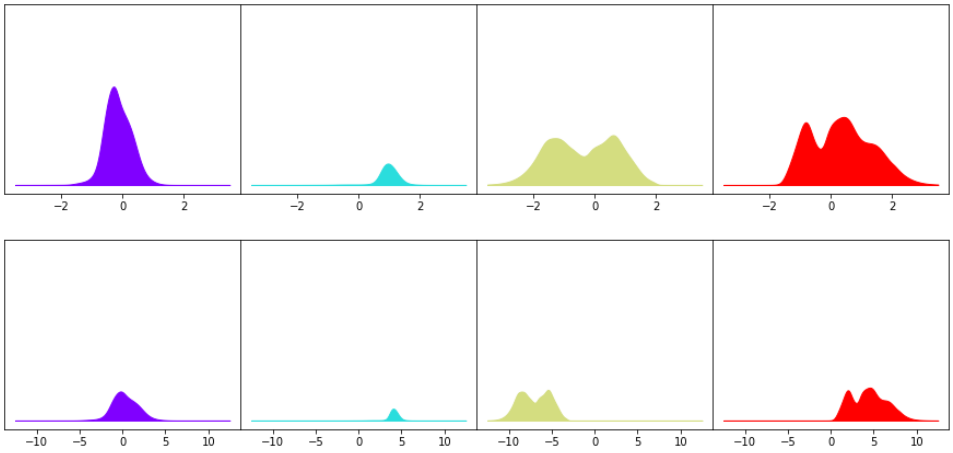
It is interesting to note that a DIF is particularly well suited for capturing multimodality and this is illustrated by the presented example. Indeed, two phenomenons add up: like in mixture models, the elements of mass are dispatched into several regions of space; but these elements themselves can be turned multimodal, since the prior q is reshaped by a function $w_k(z)$. For instance in figure 4.8, a distribution with 5 modes was captured with only $K = 4$ components.



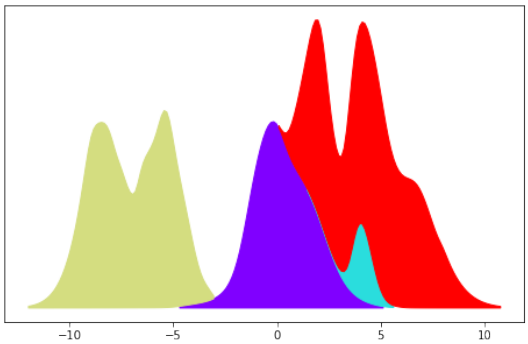
(a) A DIF transforms the prior PDF $\mathcal{Q}$ (green) into the final PDF $\mathcal{P}_\theta$ (blue)



(b) Partitioning of the latent space with functions $w_k(z)$



(c) Elements of prior mass transported with mappings T_k^{-1}



(d) Recombination

Figure 4.8: DIF mechanism for decomposing / recomposing $\mathcal{Q}$ into $\mathcal{P}_\theta$

4.3.9 Experiments

We conducted two experiments with the considered DIF model with the parameterization proposed before. First, throughout a quantitative study, we show that the DIF outperforms NF models on estimating the density associated with the 2-dimensional distribution associated with a greyscale image. This result provides empirical evidence that the proposed architecture is able to efficiently represent distributions with fine details and sharp edges in a low dimensional setting. Second, we used the DIF architecture to approximate the MNIST handwritten digit database (57) distribution in order to challenge the DIF model on a higher dimensional estimation problem.

Comparing DIF to NF architectures for capturing details in 2-dimensional portrait images

We first illustrate the strength of the DIF model on a 2-dimensional VDE problem, and provide empirical evidence that such a DIF architecture is better suited for representing distributions with fine details and sharp edges than NF or GMM models.

Let us first describe the experimental setting: we consider a greyscale (portrait of Euler) image as a 2-dimensional simple function which can thus also be seen as the PDF associated to a mixture of uniform distributions. We can easily obtain samples from this distribution by first sampling a (categorical) pixel location with probability proportional to the pixel intensity values, and then sampling a point uniformly on that pixel. With this procedure, we can obtain a set of samples from the underlying 2-dimensional distribution of an image.

Figure 4.9: A gray scale image of Euler (left) and samples from its underlying 2-dimensional distribution (right)

Using a parametric family, we can then perform VDE from the set of samples, and we here compare the performance of different parametric models in order to obtain a variational approximation of the probability distribution. We show that a DIF architecture outperforms NF architectures of Real-NVP (RNVP)(26) as well as Neural Spline

Flow (NSF)(30) (which has perhaps become the state of the art) on the task of VDE. More precisely, with less trainable parameters, we are able to reach higher log-likelihood scores with a DIF than with the two NF architectures. Moreover, as a baseline reference we also include the result of VDE using a GMM with full covariance matrix trained with the EM algorithm (24), though it has fewer parameters than the other architectures. The next table displays the log-likelihood scores of DIF, RNVP, NSF and GMM computed on 20 independent runs:

Architecture	DIF	RNVP	NSF	GMM
Parameters	39925	85780	79560	343
Log-Likelihood	$\mathbf{0.221 \pm 1.6e-3}$	$0.133 \pm 1.1e-2$	$0.182 \pm 3.9e-3$	$0.150 \pm 2.2e-3$

Table 4.1: Number of parameters and log-likelihood of different architectures for estimating the density of a 2-dimensional portrait of Euler distribution.

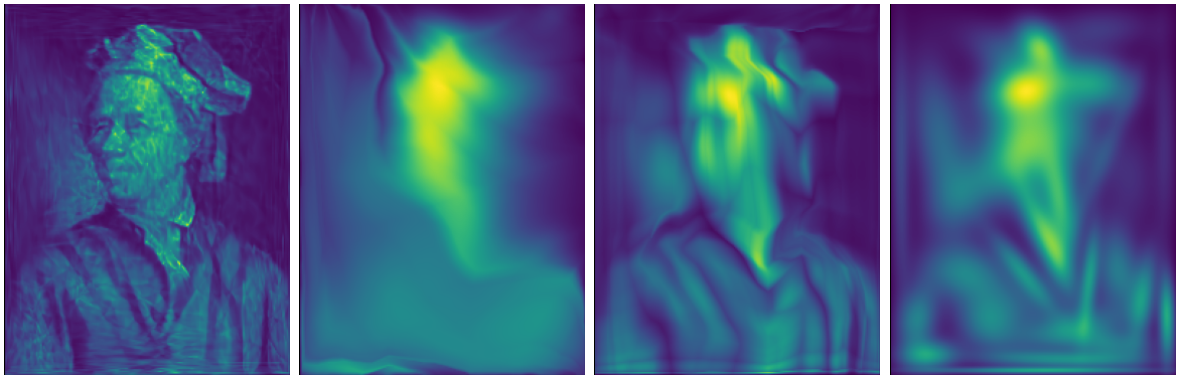


Figure 4.10: Estimated densities respectively from DIF, RNVP, NSF and GMM

For each model, we display an estimated density in figure 4.10, and we observe that the visual results are coherent with the log-likelihood scores; indeed the DIF produces a density function which most closely resembles the original image.

From this example we see that leveraging flexible probability functions $w_k(z)$ as those proposed in section 4.3.8 enables a DIF to reach distributions with sharp edges, and even close to discontinuous PDF. By contrast, we see that NFs and full covariance matrix GMM fail at efficiently capturing the finer details of the distribution.

MNIST

We then tested the proposed model on the higher dimensional problem of learning the distribution of the MNIST handwritten digit database. Here, unlike in the previous section, the 28x28 greyscale images of digit are not treated as individual 2-dimensional distributions, but as a sample from a distribution of dimension 784. The samples are the flatten images with its pixel intensities taking integer values between 0 and 255. Hence, in order to artificially make the distribution continuous, we added a uniform noise to the samples to obtain values in $\big[0, 256\big]$. Finally, we rescaled the samples between 0 and 1 by dividing throughout by 256 and we transformed the data by applying a logit

transformation of the form $x \to \text{logit}(\lambda + (1 - 2\lambda)x)$ where $\lambda = 1e - 6$ (see (26) §4.1 for more details).

In this MNIST experiment, we started off by using the proposed DIF architecture to learn this distribution. With our proposed parameterization, we obtained a log-likelihood score computed on a test set of the MNIST distribution of -1343.06 ± 0.426. It turns out that this was not a significant improvement when compared to a simple Gaussian model with estimated mean and covariance where the test log-likelihood score is -1366.22. This result is somehow expected since, with a location-scale transport, the different dimensions of a latent sample vector are modified independently from one another. Hence this transformation does not capture correlation between dimensions (i.e. the pixels).

Then, in order to solve this correlation problem, we replaced the simple location-scale with an Autoregressive transformation (35),(68) which aims at capturing the correlation between the pixels by transforming each dimension via a mapping which depends on the other dimensions as well. With this DIF parameterization (which was actually constructed as a cascade of a location-scale DIF with a Masked-Autoregressive NF - see section 4.3.10) we obtained comparable performances to an NF with approximately the same number of parameters, which is presented in the next table:

Architecture	DIF	Masked Autoregressive NF
Parameters	3189428	3357760
Train Log-Likelihood	-1220.65 ± 1.29	$\mathbf{-1217.53 \pm 1.18}$
Test Log-Likelihood	$\mathbf{-1264.72 \pm 1.78}$	-1271.12 ± 1.97

Table 4.2: Number of parameters and log-likelihood scores computed on a test set of a DIF with Autoregressive mapping versus a Masked Autoregressive NF

From this toy experiment, we can already conclude that the DIF with simple location scale transformation on its own is not capable to fully capture the correlation with the pixels with only the flexible weights $w_k(z)$. Indeed, it does not significantly outperforms a simple Gaussian model with estimated mean and covariance. On the other hand, if we select a flexible mapping T_k which is better suited to transform the dimensions dependently from one another, we are able to reach a log-likelihood score comparable to that of an NF.

Indeed, with the same number of parameters, the DIF (slightly) outperforms the NF. Finally, this experiment may indicate that the DIF is less susceptible to overfitting on the training samples than the NF. Indeed, we obtained a higher training loss but a better score on a test set.

GMM are compatible and relevant initial structures for training DIF

So far, we have presented DIF as a non-deterministic extension of deterministic NF. However, as already mentioned in section 4.3.5, due to their discrete latent structure, DIF can be connected to mixture models as well. More precisely, in section 4.3.8, we explained that with considering a convenient and practical parameterization, we obtained a model which can be seen as an extension of GMM where the mixture weights are not

necessarily constant values but can instead be represented by a K-label classification function. The purpose of this section is to explain that, in the VDE setting and with considering a location-scale parameterization as proposed in section 4.3.8, we can use a relevant GMM (which can be obtained via EM) as an initialization point for DIF training. This enables to speed up the DIF training procedure, as well as making it more numerically stable.

First, note that if the functions w_k take constant values, then the corresponding distribution is a GMM. Conversely, we can ensure that a K-label classifier function predicts constant values. This can be achieved effortlessly by simply setting to zero W_L (the weights of the last hidden layer which computes w_k), and setting b_L to the desired values.

Moreover, in the specific VDE setting, a drawback of DIF as compared to GMM is that we no longer dispose of an efficient optimization procedure. More precisely, when the weights are constant, the maximum of (4.79) can be computed in closed-form, which yields the well-known EM procedure for GMM (24). DIF no longer benefit from the same advantage, since neither the log-likelihood function (4.66) nor the surrogate function (4.79) admit closed form maxima, and we therefore can only resort to gradient-based optimisation procedures.

However, we can force a DIF to represent a given GMM and therefore we can use a specific GMM as an initialization for a DIF. We propose the idea of initializing a DIF with a GMM obtained via EM: first, train a GMM model with the EM algorithm, then set a DIF to that GMM (as explained before, by initializing the location and scale parameters as well as the probability functions w_k accordingly) and maximize the log-likelihood with a gradient based algorithm. This 2-step procedure is summarized in figure 4.11 in which we considered the same experimental setting as in section 4.3.9 on Gauss and Laplace portraits.

4.3.10 Cascading DIF

We now see that, in same spirit as NF, we can cascade simple DIF together in order to produce expressive models.

Methodology

Remember from section 4.3.3 that NF can be constructed as a composition of successive transforms. As we now see, it is also possible to cascade DIF themselves: stacking two (or more) DIF produces a DIF, so DIF can be used as elementary building blocks for defining elaborate models and transforms. To see this, consider the following cascade of two DIF $\overleftarrow{\Pi}_\theta^{[0]}$ and $\overleftarrow{\Pi}_\theta^{[1]}$:

$$\begin{array}{ccccc} x & & z_1 & & z \\ \mathcal{P} & \xrightarrow{\overrightarrow{\Pi}_\theta^{[0]}(x)} & \mathcal{Q}_\theta^{[0]} & \xrightarrow{\overrightarrow{\Pi}_\theta^{[1]}(z_1)} & \mathcal{Q}_\theta \\ \mathcal{P}_\theta & \xleftarrow{\overleftarrow{\Pi}_\theta^{[0]}(z_1)} & \mathcal{P}_\theta^{[1]} & \xleftarrow{\overleftarrow{\Pi}_\theta^{[1]}(z)} & Q \end{array}$$

Figure 4.11: A DIF (right) can approach the distribution associated to an image (left) from samples (middle-left) - with being initialized to a GMM obtained via EM (middle-right)

It is easy to see that the equivalent backward transportation is given by:

$$\overleftarrow{\pi}_\theta^{[0,1]}(x|z) = \sum_{k_1=1}^{K_1}\sum_{k_0=1}^{K_0} w_{k_0,k_1}(z)\delta_{T_{k_0,k_1}^{-1}(z)}(x), \tag{4.82}$$

$$T_{k_0,k_1}(x) = T_{k_1}^{[1]}\Big(T_{k_0}^{[0]}(x)\Big), \tag{4.83}$$

$$w_{k_0,k_1}(z) = w_{k_0}^{[0]}\Big({T_{k_1}^{[1]}}^{-1}(z)\Big)w_{k_1}^{[1]}(z). \tag{4.84}$$

Comparing (4.82) with (4.66), we see that $\overleftarrow{\Pi}_\theta^{[0,1]}$ is indeed a DIF with $K_1 \times K_0$ components; the probability mass is split into $K_1 \times K_0$ components, but by using only the equivalent of $K_1 + K_0$ parameters.

This construction generalizes the discussion in section 4.3.3 (which corresponds to the case $K_0 = K_1 = 1$), and includes as particular cases the cascading of DIF with NF ($K_0 = 1$ or $K_1 = 1$). Moreover, we see from (4.83) and (4.84) that apart from an increased number of components (which are correlated since they share a smaller set of parameters), cascading DIF potentially enables to create elaborate mappings from simple ones.

In particular, cascading a DIF ($K_1 > 1$) with simple mappings $T_{k_1}^{[1]}$ (such as location-scale) with an NF ($K_0 = 1$) in which mapping T is a flexible change of variables (such as (25) (26) (68) (52) (51) (30)(59)) produces a new DIF with K_1 components, but with more flexible mappings than those used in the initial DIF. Of course, the discussion in

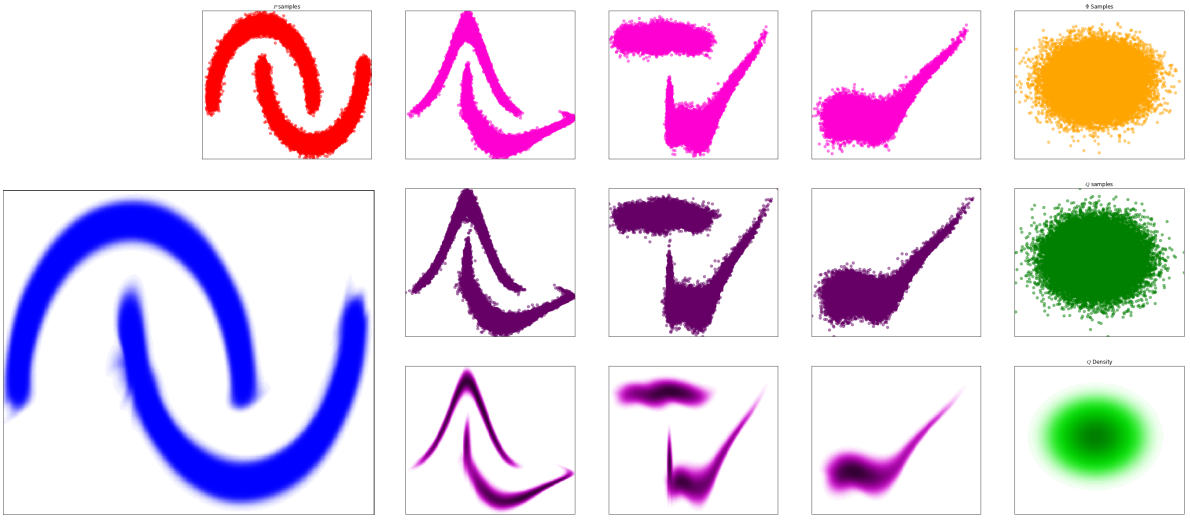


Figure 4.12: Breaking the topological barrier using a DIF within an NF

this section can be extended to more than two DIF as the principles apply recursively. Finally, cascading DIF induces a particular expression of the D_{KL} to be minimized (both for the VI and the VDE problems), see appendix A.6 for details.

Breaking topological limitation of NF

Remember from section 4.3.8 that the prior probability mass is split into several components, enabling DIF to express multimodal PDF and/or PDF with disjoint support. So DIF can be used for breaking the topological limitation evoked in section 4.3.3.

Let us illustrate this via the following example. In figure 4.12 we display how a Gaussian prior can be transformed simply into a PDF with disjoint support. This is achieved by a DIF which was built as a cascade, as explained in section 4.3.10. More precisely, the only difference with figure 4.5 is that a DIF was included in the flow steps (in between the 2nd and the 3rd). This DIF element was purposely chosen to be simple with $K = 2$ components: its only role is to separate the mass into two elements; the flexibility of the whole transform is otherwise guaranteed by the NF steps with complex deterministic mappings. The interpretation of the figure is otherwise the same.

4.3.11 Conditional Density Estimation using DIF

Up to now we have focused on the problem of modeling an unconditional probability distribution, be it for the VI or the VDE problems. However, for approximate inference purposes in the settings of classification, regression and in the likelihood-free inference (58), modeling a conditional PDF $p(x|\omega)$, where ω is some covariate rv, is also a relevant problem. It happens that NF can easily be turned into Conditional Density models; as we now see, DIF can also be used for the same purpose. In this section we will briefly explain the principles of CDE using NF, and see how the discussion can be extended to DIF.

Recall that an NF is given by a change of variables T with input z; therefore in order to obtain a conditional NF, the mapping T must be a function of the covariate ω, such that for fixed ω, $T(.;\omega)$ is a C1-diffeomorphism. This corresponds to defining a conditional transport $\pi(x|z,\omega) = \delta_{T^{-1}(z;\omega)}(x)$, so the resulting conditional PDF reads $p_\theta(x|\omega) = q\big(T(x;\omega)\big)|\det \mathrm{J}_{T(.;\omega)}(x)|$. In practice, since the mapping T is classically parameterized by an NN, this can be achieved for instance by augmenting the input of the NN with the covariate ω (see for example (68, §3.4)).

Now if we relax the hypothesis of an invertible deterministic transport and consider a discrete conditional stochastic transport of the form $\pi(x|z,\omega) = \sum_{k=1}^K w_k(z;\omega)\delta_{T_k^{-1}(z;\omega)}(x)$, then we obtain a conditional DIF model with PDF:

$$p_\theta(x|\omega) = \sum_{k=1}^K w_k\Big(T_k\big(x;\omega\big);\omega\Big) q\Big(T_k\big(x;\omega\big)\Big)\left|\det \mathrm{J}_{T_k\left(.;\omega\right)}(x)\right|$$

Therefore, in order to use a DIF as a CDE model, we simply transform T_k and w_k (for $k = 1, ..., K$) into functions of the covariate ω. For fixed ω, the associated transform is a DIF as defined above, and as such benefits from straightforward sampling and evaluation of the PDF (see section 4.3.5).

Let us propose a parameterization of conditional DIF in the spirit of section 4.3.8. First, the probability functions described in 4.3.8 can effortlessly be turned into conditional partitioning functions of the latent space z which depend on ω, denoted as $w_k(z;\omega)$. Indeed, by augmenting the input z with the covariate ω, the output of the NN is now the vectors of categorical probabilities $\Big[w_1\big(z;\omega\big), ..., w_K\big(z;\omega\big)\Big]$. Next C1-diffeomorphisms $T_1, ..., T_K$ can be turned into functions of the covariate ω by simply turning the locations and scales into functions of ω. We can use an approach similar to that used in Mixture Density Networks (MDN) (6), where an NN function predicts the location $m_k(\omega)$ and log-scales $\log\big(s_k(\omega)\big)$ for $k = 1, ..., K$.

Let us finally consider the optimization objective involved in the CDE problem (independently of the structure used for the surrogate $p_\theta(x|\omega)$, be it a DIF, an NF, an MDN or another model). We assume that we dispose of samples $\big(\omega_i, x_i\big)$ for $i = 1, ..., M$, such that $\omega_i \sim p_\omega(\omega)$ (the prior PDF of RV ω) and $x_i \sim p(x|\omega_i)$, but the conditional PDF $p(x|\omega)$ cannot be evaluated. We will build the conditional surrogate $p_\theta(x|\omega)$ of $p(x|\omega)$ by minimizing the following D_{KL}:

$$\arg\min_{\mathcal{P}_\theta} D_{\mathrm{KL}}\big(p_\omega(\omega)p(x|\omega)||p_\omega(\omega)p_\theta(x|\omega)\big) = \arg\max_{\mathcal{P}_\theta} \mathbb{E}_{p_\omega(\omega)p(x|\omega)}\Big[\log\big(p_\theta(x|\omega)\big)\Big].$$

In the end, by using an MC approximation of this expectation based on the samples at hand, the D_{KL} minimization reduces to maximizing the conditional likelihood:

$$\arg\max_{\mathcal{P}_\theta} \sum_{i=1}^M \log\big(p_\theta(x_i|\omega_i)\big).$$

4.4 Conclusion

In this work, we have explored DIF as a methodology to construct parametric surrogates in order to tackle the VI or VDE problems. As an extension of NF, DIF produce

high flexibility while remaining convenient to use as they are well suited for sampling and density estimation; moreover they do not suffer from the NF topological limitation when targeting PDF with disjoint support. On the other hand, DIF also extend mixture density models, and leverage flexible partitioning functions in order to capture detailed and edged distributions.

4.5 Perspectives and future work

4.5.1 Universal approximation with DIF ?

The DIF methodology introduces a new way to parameterize probability distributions using flexible NN functions, and we have described several advantages of this construction. Its main appeal is that this model benefits from a tractable PDF, making it an appealing tool for many tasks.

On the one hand, as we have mentioned in this chapter (with relevant references), GMMs are universal approximation tools for density functions. This means that if we increase the number of components, we can, at least in theory, reach arbitrary approximation precision. However, the more components are considered, the more numerically intensive the parameter estimation (possibly via the EM algorithm), the sampling from the corresponding model, and the PDF evaluation are. So in practice, we instead limit the number of components, and this precisely motivated the construction of DIF, where the aim is to induce further flexibility for a fixed number of components by replacing the constant mixture weights by a flexible, possibly NN-based, function. On the other hand, some NN functions are also universal approximation tools (see, for example, (21), where the authors consider increasing width and sigmoid activation functions). DIF, as an extension of mixture models, also inherits the same universal approximation property. However, an interesting perspective for future work would be to determine if, for a fixed number of components and by instead leveraging an arbitrary flexible NN partitioning function instead of the constant mixture weights, the corresponding DIF construction also provides a universal approximation of distributions.

4.5.2 Towards continous LVMs with tractable PDFs

LVMs such as VAEs, NFs and GANs have become popular tools in the machine learning community. They have yielded significant improvements compared to traditional methods, and paved the way for further developments in generative modeling techniques. They have established the potency of LVMs couples with NN and, notably, they have led to the development and popularization of several continuous LVMs such as the aforementioned continuously indexed flows (20). Among the prominent approaches are diffusion models and continuous normalizing flows. Diffusion models typically utilize stochastic differential equations to define a diffusion process that gradually transforms simple initial distributions into complex target distributions. In contrast, continuous normalizing flows employ ordinary differential equations to construct invertible transformations of probability densities. Due to the continuous nature of the latent variable,

both models lack a PDF in closed form. Nonetheless, the resulting PDFs can be approximated using sophisticated solvers for differential equations. Currently, using such solvers for PDF approximation remains costly and inefficient in practice. However, this realization opens up the possibility of fast and accurate PDF evaluation in (deep) LVMs with continuous latent variables in the future.

Bibliography

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. {TensorFlow}: a system for {Large-Scale} machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283, 2016.

[2] Maksym Andriushchenko and Nicolas Flammarion. Understanding and improving fast adversarial training. *Advances in Neural Information Processing Systems*, 33:16048–16059, 2020.

[3] Matthias Bauer and Andriy Mnih. Generalized doubly reparameterized gradient estimators. In *International Conference on Machine Learning*, pages 738–747. PMLR, 2021.

[4] Atilim Gunes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of Marchine Learning Research*, 18:1–43, 2018.

[5] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[6] Christopher M Bishop. Mixture density networks. 1994.

[7] David Blackwell. Conditional expectation and unbiased sequential estimation. *The Annals of Mathematical Statistics*, pages 105–110, 1947.

[8] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[9] Vladimir Igorevich Bogachev, Aleksandr Viktorovich Kolesnikov, and Kirill Vladimirovich Medvedev. Triangular transformations of measures. *Sbornik: Mathematics*, 196(3):309, 2005.

[10] Russell A Boyles. On the convergence of the EM algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 45(1):47–50, 1983.

[11] Tom Brijs, Dimitris Karlis, Gilbert Swinnen, Koen Vanhoof, Geert Wets, and Puneet Manchanda. A multivariate Poisson mixture model for marketing applications. *Statistica Neerlandica*, 58(3):322–348, 2004.

[12] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

[13] Craig Calcaterra. Linear Combinations of Gaussians with a Single Variance are Dense in L2. *Lecture Notes in Engineering and Computer Science*, 2171, 07 2008.

[14] O. Cappé, É. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer-Verlag, 2005.

[15] George Casella and Christian P Robert. Rao-Blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94, 1996.

[16] Anthony Caterini, Rob Cornish, Dino Sejdinovic, and Arnaud Doucet. Variational inference with continuously-indexed normalizing flows. In *Uncertainty in Artificial Intelligence*, pages 44–53. PMLR, 2021.

[17] Gilles Celeux and Gérard Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis*, 14(3):315–332, 1992.

[18] Liqun Chen, Chenyang Tao, Ruiyi Zhang, Ricardo Henao, and Lawrence Carin Duke. Variational inference and model selection with generalized evidence bounds. In *International conference on machine learning*, pages 893–902. PMLR, 2018.

[19] Siddhartha Chib and Edward Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

[20] Rob Cornish, Anthony Caterini, George Deligiannidis, and Arnaud Doucet. Relaxing bijectivity constraints with continuously indexed normalising flows. In *International conference on machine learning*, pages 2133–2143. PMLR, 2020.

[21] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[22] RUBIN DB. Using the SIR algorithm to simulate posterior distributions. In *Bayesian statistics 3. Proceedings of the third Valencia international meeting, 1-5 June 1987*, pages 395–402. Clarendon Press, 1988.

[23] Jean Pierre Delmas. An equivalence of the EM and ICE algorithm for exponential family. *IEEE transactions on signal processing*, 45(10):2613–2615, 1997.

[24] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

[25] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear Independent Components Estimation. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.

[26] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[27] Laurent Dinh, Jascha Sohl-Dickstein, Razvan Pascanu, and Hugo Larochelle. A RAD approach to deep mixture models. In *Deep Generative Models for Highly Structured Data, ICLR 2019 Workshop, New Orleans, Louisiana, United States, May 6, 2019*. OpenReview.net, 2019.

[28] Arnaud Doucet, Eric Moulines, and Achille Thin. Differentiable samplers for deep latent variable models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2247), March 2023.

[29] Leo L. Duan. Transport Monte Carlo: High-Accuracy Posterior Approximation via Random Transport. *Journal of the American Statistical Association*, 0(0):1–12, 2023.

[30] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in Neural Information Processing Systems*, 32:7511–7522, 2019.

[31] Jeffrey A Fessler and Alfred O Hero. Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithms. *IEEE Transactions on Image Processing*, 4(10):1417–1429, 1995.

[32] Michael Figurnov, Shakir Mohamed, and Andriy Mnih. Implicit reparameterization gradients. *Advances in Neural Information Processing Systems*, 31, 2018.

[33] Yarin Gal et al. Uncertainty in deep learning. 2016.

[34] A. E. Gelfand and A. F. M. Smith. Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

[35] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. Made: Masked autoencoder for distribution estimation. In *International conference on machine learning*, pages 881–889. PMLR, 2015.

[36] Paul Glasserman. *Monte Carlo methods in financial engineering*, volume 53. Springer, 2004.

[37] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[38] Alex Graves. Stochastic backpropagation through mixture density distributions. *arXiv preprint arXiv:1607.05690*, 2016.

[39] Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free MCMC with Amortized Approximate Ratio Estimators. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4239–4248. PMLR, 13–18 Jul 2020.

[40] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[41] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(5), 2013.

[42] Chin-Wei Huang, Laurent Dinh, and Aaron Courville. Augmented normalizing flows: Bridging the gap between generative flows and latent variable models. *arXiv preprint arXiv:2002.07101*, 2020.

[43] David R Hunter and Kenneth Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37, 2004.

[44] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

[45] Martin Jankowiak and Fritz Obermeyer. Pathwise derivatives beyond the reparameterization trick. In *International conference on machine learning*, pages 2235–2244. PMLR, 2018.

[46] Thomas Kämpke, Franz Josef Radermacher, Thomas Kämpke, and Franz Josef Radermacher. The Generalized Inverse of Distribution Functions. *Income Modeling and Balancing: A Rigorous Treatment of Distribution Patterns*, pages 9–28, 2015.

[47] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[48] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[49] Diederik P. Kingma and Jimmy Ba. ADAM: A Method for Stochastic Optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[50] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[51] Durk P Kingma and Prafulla Dhariwal. Glow: Generative Flow with Invertible 1x1 Convolutions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[52] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29, 2016.

[53] Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015.

[54] Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker. Normalizing Flows: An Introduction and Review of Current Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, 2021.

[55] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *Journal of machine learning research*, 2017.

[56] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[57] Yann LeCun and Corinna Cortes.

[58] Jan-Matthis Lueckmann, Giacomo Bassetto, Theofanis Karaletsos, and Jakob H Macke. Likelihood-free inference with emulator networks. In *Symposium on Advances in Approximate Bayesian Inference*, pages 32–53. PMLR, 2019.

[59] Chenlin Meng, Linqi Zhou, Kristy Choi, Tri Dao, and Stefano Ermon. Butterflyflow: Building invertible layers with butterfly matrices. In *International Conference on Machine Learning*, pages 15360–15375. PMLR, 2022.

[60] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018.

[61] Warren Morningstar, Sharad Vikram, Cusuh Ham, Andrew Gallagher, and Joshua Dillon. Automatic differentiation variational inference with mixtures. In *International Conference on Artificial Intelligence and Statistics*, pages 3250–3258. PMLR, 2021.

[62] Christian Naesseth, Francisco Ruiz, Scott Linderman, and David Blei. Reparameterization gradients through acceptance-rejection sampling algorithms. In *Artificial Intelligence and Statistics*, pages 489–498. PMLR, 2017.

[63] Didrik Nielsen, Priyank Jaini, Emiel Hoogeboom, Ole Winther, and Max Welling. Survae flows: Surjections to bridge the gap between vaes and flows. *Advances in Neural Information Processing Systems*, 33:12685–12696, 2020.

[64] Robert D Nowak. Distributed EM algorithms for density estimation and clustering in sensor networks. *IEEE transactions on signal processing*, 51(8):2245–2253, 2003.

[65] Manfred Opper and Cédric Archambeau. The variational Gaussian approximation revisited. *Neural computation*, 21(3):786–792, 2009.

[66] John Paisley, David Blei, and Michael Jordan. Variational Bayesian inference with stochastic search. *arXiv preprint arXiv:1206.6430*, 2012.

[67] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):2617–2680, 2021.

[68] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30, 2017.

[69] Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

[70] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[71] Kostantinos N Plataniotis and Dimitris Hatzinakos. Gaussian mixtures and their applications to signal processing. *Advanced signal processing handbook*, pages 89–124, 2017.

[72] Tom Rainforth, Adam Kosiorek, Tuan Anh Le, Chris Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter variational bounds are not necessarily better. In *International Conference on Machine Learning*, pages 4277–4285. PMLR, 2018.

[73] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[74] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

[75] Christian P. Robert and Gareth Roberts. Rao–Blackwellisation in the Markov Chain Monte Carlo Era. *International Statistical Review*, 89(2):237–249, 2021.

[76] Francisco R Ruiz, Titsias RC AUEB, David Blei, et al. The generalized reparameterization gradient. *Advances in neural information processing systems*, 29, 2016.

[77] John Schulman, Nicolas Heess, Theophane Weber, and Pieter Abbeel. Gradient estimation using stochastic computation graphs. *Advances in neural information processing systems*, 28, 2015.

[78] A. F. M. Smith and A. E. Gelfand. Bayesian statistics without tears : a sampling-resampling perspective. *The American Statistician*, 46(2):84–87, 1992.

[79] Adrian FM Smith and Gareth O Roberts. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(1):3–23, 1993.

[80] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[81] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. *Advances in neural information processing systems*, 29, 2016.

[82] Ying Sun, Prabhu Babu, and Daniel P Palomar. Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Transactions on Signal Processing*, 65(3):794–816, 2016.

[83] Achille Thin, Nikita Kotelevskii, Jean-Stanislas Denain, Leo Grinsztajn, Alain Durmus, Maxim Panov, and Eric Moulines. MetFlow: A New Efficient Method for Bridging the Gap between Markov Chain Monte Carlo and Variational Inference, 2020.

[84] Michalis Titsias and Miguel Lázaro-Gredilla. Doubly stochastic variational Bayes for non-conjugate inference. In *International conference on machine learning*, pages 1971–1979. PMLR, 2014.

[85] Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.

[86] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

[87] CF Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of statistics*, pages 95–103, 1983.

[88] Ming Xu, Matias Quiroz, Robert Kohn, and Scott A Sisson. Variance reduction properties of the reparameterization trick. In *The 22nd international conference on artificial intelligence and statistics*, pages 2711–2720. PMLR, 2019.

[89] David Zoltowski, Diana Cai, and Ryan P Adams. Slice Sampling Reparameterization Gradients. *Advances in Neural Information Processing Systems*, 34:23532–23544, 2021.

Conclusion

Dans le paysage technologique en rapide évolution d'aujourd'hui, l'apprentissage automatique et l'apprentissage statistique jouent des rôles essentiels dans la transformation des industries et des services, et dans la redéfinition des expériences quotidiennes. Ces méthodes utilisent des modèles sophistiqués et permettent d'apprendre à partir des données enregistrées pour faire des prédictions. Facilitée par les améliorations de la puissance de calcul et des modèles mathématiques, l'apprentissage automatique stimule les innovations dans divers secteurs, de la santé et des finances au marketing et au divertissement. À l'ère des mégadonnées et de la nouvelle capacité de calcul, les algorithmes d'apprentissage automatique peuvent découvrir des insights et des tendances que les méthodes statistiques traditionnelles pourraient négliger. Cette capacité à traiter efficacement d'immenses quantités de données ouvre la voie à une révolution dans des domaines tels que la médecine personnalisée, les systèmes de recommandation et les véhicules autonomes.

De plus, les techniques d'apprentissage automatique telles que le deep learning ont considérablement amélioré les performances de tâches comme la reconnaissance d'images et de la parole, le traitement du langage naturel, et même les jeux vidéo. Ces avancées soulignent le potentiel de l'apprentissage automatique pour créer des systèmes plus intelligents capables de s'adapter et de s'améliorer au fil du temps. Alors que nous naviguons dans cette ère de transformation, il devient de plus en plus crucial de comprendre les principes et les applications de l'apprentissage automatique et de l'apprentissage statistique. Ces technologies ne conduisent pas seulement à l'innovation et à l'efficacité, mais soulèvent également des questions éthiques et sociétales importantes concernant la confidentialité, les biais et la transparence. Par conséquent, une approche équilibrée de leur développement et de leur déploiement est essentielle pour exploiter pleinement leur potentiel tout en abordant ces défis potentiels.

Les méthodes d'apprentissage automatique probabiliste, en tant qu'extension des principes bayésiens, offrent un puissant paradigme pour développer des modèles qui sont non seulement prédictifs mais aussi intrinsèquement interprétables. En adoptant la perspective probabiliste, ces modèles fournissent des insights sur les processus générateurs de données sous-jacents et nous permettent de saisir les nuances des phénomènes du monde réel dans les tâches de classification et de régression. La synthèse de modèles probabilistes avec des algorithmes d'apprentissage automatique a conduit au développement de techniques innovantes qui repoussent les limites de ce qui est possible dans la recherche basée sur les données. Contrairement aux approches traditionnelles qui fournissent des résultats déterministes, les méthodes probabilistes attribuent des

probabilités aux résultats, offrant une compréhension plus riche de l'incertitude et de la variabilité des données. Cela permet une prise de décision plus nuancée, une gestion robuste des données incomplètes, et la capacité de quantifier et de gérer les risques de manière efficace, ce qui les rend particulièrement précieuses dans les applications où prendre des décisions éclairées est crucial. Cependant, elles présentent leurs propres défis, tels que l'évolutivité, la sélection des modèles et, bien sûr, le coût computationnel.

L'inférence bayésienne, avec son cadre robuste pour la mise à jour des croyances à la lumière de nouvelles preuves, s'est avérée être un pilier fondamental du raisonnement probabiliste. En utilisant les méthodes bayésiennes, il est possible d'appliquer les observations du monde réel à des modèles scientifiques pertinents, d'incorporer des connaissances antérieures, de quantifier l'incertitude aléatoire, et de prendre des décisions éclairées même dans des environnements complexes et incertains. Les mises en œuvre pratiques de l'inférence bayésienne dans divers domaines, de la traitement du langage naturel à l'ingénierie biomédicale, soulignent sa polyvalence et son efficacité. Dans des domaines tels que la génétique, les sciences de l'environnement et la médecine, les méthodes bayésiennes permettent aux chercheurs de prendre des décisions éclairées face à l'incertitude et de prédire les résultats futurs avec une plus grande confiance. Dans les contextes de prise de décision, de la finance aux politiques publiques, l'inférence bayésienne soutient une évaluation robuste des risques et une planification stratégique en affinant continuellement les prévisions au fur et à mesure que de nouvelles informations deviennent disponibles. Malgré ses défis computationnels, notamment dans les espaces de haute dimension, les avancées dans les algorithmes et la puissance de calcul, telles que les méthodes de Monte Carlo par chaîne de Markov et l'Inference Variationnelle, ont rendu l'inférence bayésienne plus tractable et efficace. En conséquence, les méthodes bayésiennes continuent de stimuler l'innovation, soulignant leur importance durable et leur potentiel pour des percées futures tant dans la découverte scientifique que dans la prise de décision pratique.

Les principes de l'inférence bayésienne postérieure peuvent donc être utilisés pour étudier, interpréter, analyser et modéliser des phénomènes réels inconnus à partir des observations dont nous disposons. Cela peut être réalisé en sélectionnant une variable explicative appropriée, en obtenant des informations antérieures sur cette quantité et en choisissant un modèle d'observation. Cependant, lorsqu'on considère des modèles scientifiques élaborés, nous sommes souvent confrontés à un problème d'inférence bayésienne postérieure difficile à résoudre. Cela se produit lorsque la relation entre la variable observée et la variable explicative est décrite par un modèle d'observation dont la distribution sous-jacente ne bénéficie pas d'une fonction de densité de probabilité (PDF) tractable (appelée cadre sans vraisemblance), comme lorsqu'elle est conçue via un modèle de simulation implicite, rendant également la PDF postérieure inaccessible à l'évaluation. Il est toutefois possible que nous disposions d'observations enregistrées à partir du modèle d'observation, auquel cas nous pouvons recourir à une inférence postérieure approximative en utilisant des méthodes d'apprentissage statistique.

L'objectif de cette thèse est de combiner les techniques d'apprentissage automatique et d'inférence bayésienne pour réaliser une inférence basée sur un modèle approximatif d'une distribution postérieure d'intérêt. Notre but est d'apprendre un modèle statistique pour une distribution postérieure inconnue à partir d'un ensemble d'observations

générées par le modèle d'observation correspondant, dont la vraisemblance est incalculable. Cela trouve une application directe dans le cadre de l'inférence sans vraisemblance, mais fournit également une formulation probabiliste des tâches habituelles d'apprentissage statistique en classification et régression.

Comme expliqué dans le chapitre 1 de cette thèse, ce problème général soulève plusieurs questions secondaires, telles que la comparaison de différentes techniques de modélisation, l'échantillonnage à partir d'approximations postérieures, la quantification de l'incertitude épistémique, la modélisation générative etl'estimation de densité. Dans les chapitres suivants, nous avons présenté trois contributions qui répondent à certaines de ces questions. Chaque contribution peut être considérée indépendamment des autres, et nous avons fourni des détails concernant leurs contextes respectifs et les travaux connexes. Cependant, dans chaque chapitre, nous avons également cherché à relier chaque contribution au thème général de cette thèse. Ainsi, considérées ensemble, elles constituent un récit qui a été décrit dans la section d'introduction de ce document.

Pour résumer brièvement le travail de cette thèse et compléter les conclusions des chapitres individuels, nous nous référons maintenant à ce fil narratif introductif et nous décrivons maintenant les conclusions associées.

- **"Le *likelihood-to-evidence ratio* (LTER) peut être échantillonné en utilisant les algorithmes d'échantillonnage usuels basés sur les ratio de PDF ..."**

  Dans le chapitre 2 de cette thèse, nous avons proposé la méthodologie de ≪Binary Classification Monte Carlo sampling≫. Dans cette approche, nous remplaçons le ratio de densité de probabilité dans les algorithmes d'échantillonnage usuels tels que l'acceptation-rejet, l'échantillonnage d'importance et le Metropolis-Hastings indépendant, par une approximation basée sur un classificateur. Cela transforme un problème d'échantillonnage à partir d'une distribution cible en un problème de classification, et conduit à des approches d'échantillonnage sans paramètres et sans densité, permettant ainsi l'utilisation de distributions instrumentales implicites. Cette approche peut facilement être appliquée à la technique d'approximation postérieure du rapport vraisemblance-sur-évidence (LTER).

  "... mais la question de la quantification d'incertitude dans cette modélisation reste ouverte."

  En raison de sa construction non normalisée, le LTER n'est pas facilement compatible avec la tâche de quantification de l'incertitude (UQ) via des méthodes bayésiennes. Ce problème pourrait faire l'objet de travaux futurs.

- **"La construction générative est également un modèle non-normalisée mais elle est effectivement compatible avec la quantification d'incertitude à partir de l'échantillonnage de la distribution prédictive postérieure (PPD) ..."**

Dans le chapitre 3, nous avons considéré l'approche de modélisation générative. Celle-ci consiste à modéliser la PDF du modèle d'observation inconnu à l'aide d'une loi paramétrique conditionnelle, fournissant ainsi également un modèle non-normalisé de la postérieure. Néanmoins, il s'avère que cette construction spécifique est bel et bien compatible avec la tâche de quantification de l'incertitude en utilisation la PPD, qui peut être réalisée grâce à un schéma d'échantillonnage spécifique que nous avons proposé dans ce chapitre.

"... et peut donc être comparée à la modélisation discriminante."

Les modèles discriminatifs se distinguent des modèles génératifs en ce sens que les premiers approximent directement la distribution d'intérêt, à savoir l'a posteriori, tandis que les seconds le font indirectement en approximant la fonction de vraisemblance. Cependant, les deux exploitent un modèle de distribution conditionnelle, et nous avons donc comparé les deux approches dans le cadre de l'apprentissage bayésien tenant compte de l'incertitude, via la distribution prédictive postérieure (PPD).

- **"Dans les deux approches, la quantification épistémique de l'incertitude via la PPD repose sur un modèle conditionnel avec une fonction de densité de probabilité calculable, ..."**

  Une hypothèse cruciale permettant d'utiliser la PPD dans les approches génératives et discriminatives est l'utilisation d'un modèle avec une PDF facilement calculable. Cette hypothèse garantit en effet que nous pouvons évaluer la PDF jointe sur les labels et les paramètres (ou éventuellement ses conditionnelles), étant donné les observations labelisées et/non non-labelisées. Cela permet ainsi l'échantillonnage à partir de la distribution conjointe (possiblement via un échantillonnage séquentiel des conditionnelles dans le cadre d'un schéma de Gibbs), aboutissant à des échantillons tirés de la PPD.

  "... qui peut être construit avec un Discretely Indexed Flow (DIF)."

  Nous avons proposé la construction DIF qui est une extension des modèles de mélange dans laquelle la pondération des composantes ne se fait pas en utilisant des poids scalaires mais une fonction définie par le biais d'un classifeur. L'avantage principal de ce modèle paramétrique est qu'il dispose d'une PDF facilement calculable grâce à une construction de variables latentes spécifique, tout en bénéficiant d'une flexibilité accrue par rapport aux modèles de mélanges en utilisant des fonctions de réseaux de neurones. Ses autres propriétés pratiques sont (i) un schéma d'échantillonnage simple, (ii) une reparamétrisation des gradients dans les problèmes variationnels, et (iii) la possibilité de le transformer facilement en

un modèle conditionnel. Les DIF peuvent également être considérés comme une extension des Normalizing Flows, où le mapping déterministe entre un aléatoire latent et la variable observée est remplacé par un transport stochastique de nature discrète. En tant que tel, il peut facilement être combiné avec des couches inversibles, aboutissant à un modèle mixte qui n'est plus limité par les contraintes topologiques des NFs.

Conclusion

In today's rapidly advancing technological landscape, machine learning and statistical learning play pivotal roles in transforming industries and services, and for reshaping everyday experiences. These methods use sophisticated models and enable to learn from recorded data and make predictions. Facilitated by improvements in computational power and mathematical models, machine learning is driving innovations across various sectors, from healthcare and finance to marketing and entertainment. In the era of big data and novel computational capacity, machine learning algorithms can uncover insights and trends that traditional statistical methods might overlook. This ability to process vast amounts of data efficiently paves the way for a revolution in fields like personalized medicine, recommendation systems, and autonomous vehicles.

Moreover, machine learning techniques such as deep learning have significantly enhanced the performance of tasks like image and speech recognition, natural language processing, and even game playing. These advancements underscore the potential of machine learning in creating smarter systems capable of adapting and improving over time. As we navigate this era of transformation, understanding the principles and applications of machine learning and statistical learning becomes increasingly crucial. These technologies not only drive innovation and efficiency but also raise important ethical and societal questions regarding privacy, bias, and transparency. Therefore, a balanced approach to their development and deployment is essential for leveraging their full potential while addressing such potential challenges.

Probabilistic machine learning methods, as an extension of Bayesian principles, offer a powerful paradigm for developing models that are not only predictive but also inherently interpretable. By embracing the probabilistic perspective, these models provide insights into the underlying data-generating processes and enable us to capture the nuances of real-world phenomena in classification and regression tasks. The synthesis of probabilistic models with machine learning algorithms has led to the development of innovative techniques that push the boundaries of what is possible in data-driven research. Unlike traditional approaches that provide deterministic outputs, probabilistic methods assign probabilities to outcomes, offering a richer understanding of uncertainty and variability in data. As such, it enables more nuanced decision-making, robust handling of incomplete data, and the ability to quantify and manage risks effectively, making them particularly valuable in applications where making informed decisions is critical. It comes, however, with its own challenges, such as scalability, model selection, and, of course, the computational cost.

Bayesian inference, with its robust framework for updating beliefs in light of new evi-

dence, has proven to be a cornerstone for probabilistic reasoning. By employing Bayesian methods, one can confront real-world observations to relevant scientific models, incorporate prior knowledge, quantify aleatoric uncertainty, and make informed decisions even in the face of complex and uncertain environments. The practical implementations of Bayesian inference in various domains, from natural language processing to biomedical engineering, underscore its versatility and efficacy. In fields such as genetics, environmental science, and medicine, Bayesian methods enable researchers to make informed decisions under uncertainty, and predict future outcomes with greater confidence. In decision-making contexts, from finance to public policy, Bayesian inference supports robust risk assessment and strategic planning by continuously refining predictions as new information becomes available. Despite its computational challenges, particularly in high-dimensional spaces, advancements in algorithms and computing power, such as Markov Chain Monte Carlo methods and Variational Inference, have made Bayesian inference more tractable and efficient. As a result, Bayesian methods continue to drive innovations, underscoring their enduring significance and potential for future breakthroughs in both scientific discovery and practical decision-making.

The principles of Bayesian posterior inference therefore can be used to study, interpret, analyze, and model unknown real-world phenomena from nature but from which we dispose of observations. This can be done by selecting an appropriate explanatory variable, obtaining prior information about that quantity, and choosing an observation model. However, when considering elaborate scientific models, we are often faced with an intractable Bayesian posterior inference problem. This happens when the relationship between the observed and explanatory variable is described via an observation model where the underlying distribution does not benefit from a tractable probability density function (PDF) (referred to as a likelihood-free setting), such as when it is designed via an implicit simulation model, making the posterior PDF also unavailable for evaluation. It is possible, however, that we dispose of recorded observations from the observation model, in which case we can resort to approximate posterior inference using statistical learning methods.

The scope of this thesis is to bring together machine learning and Bayesian inference to perform such a Bayesian inference based on an approximate model of a posterior distribution of interest. Our aim is to learn a statistical model for an unknown posterior distribution from recorded observations generated by the corresponding observation model with unavailable likelihood. This finds direct application in the likelihood-free inference setting, but this also provides with a probabilistic formulation of usual classification and regression statistical learning tasks.

As explained in chapter 1 of this thesis, this general problem raises several subsidiary questions, such as comparing different modeling techniques, sampling from posterior approximations, quantifying epistemic uncertainty, generative modeling, and density estimation. In the following three chapters, we provided three contributions, which answer some of these points. Each contribution can be considered independently of the others, and we provided details concerning their respective contexts and related work. However, in each chapter, we also aimed to relate each contribution to each other to the general topic of this thesis. So indeed, when considered together, they constitute a narrative which was described in the introductory section of this document.

In order to briefly summarize the findings of this thesis and to complement the conclusions of the individual chapters, we now refer back to this introductory narrative thread and provide its concluding counterpart.

- **"The likelihood-to-evidence ratio (LTER) can be sampled from using the usual ratio-based sampling algorithms; ..."**

  In chapter 2 of this thesis, we proposed the "Binary classification Monte Carlo sampling" methodology. This approach replaces the PDF ratio in the usual sampling algorithms of accept-reject, importance sampling, and independent Metropolis-Hastings, with a classifier-based approximation. This turns a problem of sampling from a target distribution into a problem of sampling classification and results in parameter-free and density-free sampling approaches, thus enabling the use of implicit instrumental distributions. This approach can easily be applied to the LTER posterior approximation technique.

  "... however it remains unclear how to apply this model in a Bayesian uncertainty-aware inference."

  Due to its unnormalized construction, the LTER is not easily compatible with the task of uncertainty quantification (UQ) using Bayesian methods. This problem can be the topic of future work.

- **"The unnormalized generative construction is indeed compatible with sampling from the posterior predictive distribution (PPD) ..."**

  In chapter 3, we considered the generative modeling approach. It consists in modeling the PDF of the unknown observation model and thus also provides an unnormalized model of the posterior. Nonetheless, it turns out that this specific construction is indeed compatible with the task of epistemic UQ via the PPD, which can be conducted via a specific sampling scheme which we proposed in this chapter.

  "... and can thus be compared to discriminative models."

  Discriminative models differ from the generative ones in the sense that the former approximates the PDF of interest, the a posteriori, directly, while the latter does so indirectly by approximating the likelihood function. However, both leverage a conditional distribution model, and we thus compared the two approaches under the scope of Bayesian uncertainty-aware learning via the PPD.

- **"In both approaches, epistemic UQ via the PPD relies on a model with**

tractable PDF, ...”

A crucial assumption which enables us to use the PPD in both generative and discriminative approaches is that of using a model with a tractable PDF. This assumption indeed ensures that we can evaluate the joint PDF over label and parameters (or possibly its conditionals), given the observations and the dataset, and thus enables the sampling from the joint distribution (possibly via sequentially sampling from the conditionals in a Gibbs sampling scheme), resulting in samples drawn from the PPD.

“... which can be constructed using Discretely Indexed Flows (DIF).”

We proposed the novel construction of DIF. The main advantage of this parametric model is that it disposes of a tractable PDF using a discrete latent variable construction, while benefiting from increased flexibility as compared to mixture models with leveraging neural functions. Its other convenient properties are straightforward (i) sampling scheme and (ii) reparameterization of gradients, and (iii) that it can easily be turned into a conditional model. DIF can also be seen as an extension of normalizing flows (NFs) where the deterministic mapping between latent noise and observed variable is replaced by a stochastic transport of discrete nature. As such, it can easily be combined (cascaded) with invertible layers, resulting in a mixed model which is no longer restricted by the topological limitations of NFs.

Appendix A

Appendix

A.1 Acronyms

ANF: Augmented Normalizing Flows
AR: Accept Reject
CDF: Cumulative Distribution Function
CIF: Continuously Indexed Flows
DE: Density Estimation
DIF: Discretely Indexed Flows
EBM: Energy Based Model
ELBO: Evidence Lower Bound
EM: Expectation-Maximization
GA: Gradient Ascent
GD: Gradient Descent
GMM: Gaussian Mixture Model
iid: Independent and Identically Distributed
IMH: Independent Metropolis Hastings
IS: Importance Sampling
LTER: Likelihood-to-evidence ratio
LVM: Latent Variable Model
MC: Monte Carlo
MCMC: Markov Chain Monte Carlo
MH: Metropolis-Hastings
MM: Majorize-Minorize
NF: Normalizing Flows
NN: Neural-Network
PDF: Probability Density Function
PPD: Posterior Predictive Distribution
RB: Rao-Blackwell
RV: Random Variable
SIR: Sampling Importance Resampling
SMC: Sequential Monte Carlo

TMC: Transport Monte Carlo
UQ: Uncertainty Quantification
VI: Variational Inference

A.2 Main notations

X: unobserved continuous or categorical random variable of interest, also refered to as *label*;
Y: observed random variable;
θ: model parameters;
Pr: probability;
p_θ: PDF associated with parametric model for given value of θ in the context of generative and discriminative modeling and DIF;
$\mathcal{Q}, q$: instrumental distribution and corresponding PDF;
$\sim$: "a random variable is distributed according to" or "a value is drawn at random from";
$\mathcal{D}$: dataset;
$\mathcal{Y}$: unlabeled observations $\{\widetilde{y}_j\}_{j=1,\dots,|\mathcal{Y}|}$ in semi-supervised learning;
$\Pi_{\_}, \pi_{\_}$: prior distribution and corresponding PDF;
D_{KL}: Kullback-Leibler divergence.

Differences throughout the chapters

In chapters 1 and 2, the prior distribution over label X is denoted $\mathcal{P}_X$ and its PDF is $p_X(x)$, while in chapter 3, prior distributions are refered to as $\pi_{\_}$ is a prior PDF (usually over RV X or θ) associated with a probability distribution $\Pi_{\_}$. In the fourth chapter, when discussing mixture models and its extension to DIF π_k denotes the mixtures weights which indeed become $\pi_k(x)$ a function of x and the conditions on $\pi_k(x)$ to result in a valid PDF are discussed. In the paper "Discretely Indexed Flows" (section 4.3), we use the notations $\overleftarrow{\Pi_\theta}$ and $\overleftarrow{\pi_\theta}$ (resp. $\overrightarrow{\Pi_\theta}$ and $\overrightarrow{\pi_\theta}$) to describe the discrete probability distribution and its mass function of observed X (resp. latent Z) given latent Z (resp. observed X).

A.3 Classifier based posterior sampling algorithms

Algorithm 5 Classifer based Importance Sampling expectation estimation

Require: observed y, function f, N, prior $\mathcal{P}_X$, classifier r_θ
  Draw samples $x_1, ..., x_N \overset{iid}{\sim} \mathcal{P}_x$
  Compute and normalize importance weights $w^{(i)} = \frac{\frac{r_\theta(x_i,y)}{1-r_\theta(x_i,y)}}{\sum_{j=1}^{N} \frac{r_\theta(x_j,y)}{1-r_\theta(x_j,y)}}$
  Compute estimate $\sum_{i=1}^{M} w^{(i)} f(x_i)$

Algorithm 6 Classifer based Sampling - Importance Resampling from the posterior

Require: observed y, N, prior $\mathcal{P}_X$, classifier r_θ
  Draw samples $x_1, ..., x_N \overset{iid}{\sim} \mathcal{P}_x$
  Compute and normalize importance weights $w^{(i)} = \frac{\frac{r_\theta(x_i,y)}{1-r_\theta(x_i,y)}}{\sum_{j=1}^{N} \frac{r_\theta(x_j,y)}{1-r_\theta(x_j,y)}}$
  while number of samples is not reached **do**
    Sample index $k \sim$ Categorical$(w^{(1)}, ..., w^{(N)})$
    Select x_k as a sample
  end while

Algorithm 7 Classifer based Independent Metropolis-Hastings MCMC

Require: observed y, T, prior $\mathcal{P}_X$, classifier r_θ
  Draw $x_0 \sim \mathcal{P}_x$
  for each Markov transition step t up to T **do**
    propose a candidate $x^* \sim \mathcal{P}_X$
    if $u \sim \mathcal{U}_{[0,1]} \leq \min(1, \frac{r_\theta(x^*,y)}{1-r_\theta(x^*,y)} \frac{1-r_\theta(x_{t-1},y)}{r_\theta(x_{t-1},y)})$ **then**
      set $x_t = x^*$
    else
      set $x_t = x_{t-1}$
    end if
  end for

Algorithm 8 Classifer based Independent Barker MCMC

Require: observed y, T, prior $\mathcal{P}_X$, classifier r_θ
  Draw $x_0 \sim \mathcal{P}_x$
  for each Markov transition step t up to T **do**
    propose a candidate $x^* \sim \mathcal{P}_X$
    if $u \sim \mathcal{U}_{[0,1]} \leq \frac{r_\theta(x^*,y)(1-r_\theta(x_{t-1},y))}{r_\theta(x^*,y)(1-r_\theta(x_{t-1},y))+r_\theta(x_{t-1},y)(1-r_\theta(x_{t-1},y))}$ **then**
      set $x_t = x^*$
    else
      set $x_t = x_{t-1}$
    end if
  end for

A.4 DIF reverse kernel and marginal distribution

We first check that the function

$$\begin{aligned}\overleftarrow{\Pi} :&\mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0,\infty[\\ &z, A \to \overleftarrow{\Pi}(z, A) = \sum_{k=1}^{K} w_k(z)\mathbb{1}(T_k^{-1}(z) \in A)\end{aligned}$$

is a valid transition kernel: for fixed $z \in \mathbb{R}^d$, $\overleftarrow{\Pi}(z, .)$ is a probability measure; while for fixed $A \in \mathcal{B}(\mathbb{R}^d)$, $\overleftarrow{\Pi}(., A)$ is a measurable function. On the other hand, $\mathbb{R}^d$ is a Polish space endowed with its Borel σ-field, so the reverse transition Kernel $\overrightarrow{\Pi}$ exists. Since for a given x, only the values $\big(T_1(x), ..., T_K(x)\big)$ may have produced x, the reverse transition kernel indeed takes the form:

$$\begin{aligned}\overrightarrow{\Pi} :&\mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0,\infty[\\ &x, B \to \overrightarrow{\Pi}(x, B) = \sum_{k=1}^{K} v_k(x)\mathbb{1}\big(T_k(x) \in B\big),\end{aligned}$$

in which $v_k(x) = \Pr(Z = T_k^{-1}(\widetilde{X})|\widetilde{X} = x)$. Next, for any $A, B \in \mathcal{B}(\mathbb{R}^d)$, we have

$$\begin{aligned}\Pr(\widetilde{X} \in A, Z \in B) &= \int_B \overleftarrow{\Pi}(z, A)q(z)\mathrm{d}z\\ &= \int_B \left(\int_A \sum_{k=1}^{K} w_k(z)\delta_{T_k^{-1}(z)}(x)\mathrm{d}x\right) q(z)\mathrm{d}z\\ &= \sum_{k=1}^{K} \int_{B\cap T_k(A)} w_k(z)q(z)\mathrm{d}z\\ &= \sum_{k=1}^{K} \int_{T_k^{-1}(B)\cap A} w_k\big(T_k(x)\big)q\big(T_k(x)\big)\big|\det \mathrm{J}_{T_k}(x)\big|\mathrm{d}x. \qquad \text{(A.1)}\end{aligned}$$

Two cases are of particular interest:

- Set $B = \mathbb{R}^d$. Since $T_k^{-1}(B) = \mathbb{R}^d$, (A.1) becomes

$$\begin{aligned}\Pr(\widetilde{X} \in A) &= \sum_{k=1}^{K} \int_A w_k\big(T_k(x)\big)q\big(T_k(x)\big)\big|\det \mathrm{J}_{T_k}(x)\big|\mathrm{d}x\\ &= \int_A \underbrace{\sum_{k=1}^{K} w_k\big(T_k(x)\big)q\big(T_k(x)\big)\big|\det \mathrm{J}_{T_k}(x)\big|}_{\psi(x)}\,\mathrm{d}x,\end{aligned}$$

  so $\widetilde{X}$ admits pdf ψ wrt Lebesgue measure.

- Set $A = \mathbb{R}^d$. Equation (A.1) becomes

$$\begin{aligned}
\Pr(Z \in B) &= \sum_{k=1}^{K} \int_{T_k^{-1}(B)} w_k\big(T_k(x)\big) q\big(T_k(x)\big) \big|\det \mathrm{J}_{T_k}(x)\big| \mathrm{d}x \\
&= \sum_{k=1}^{K} \int_{T_k^{-1}(B)} \frac{w_k\big(T_k(x)\big) q\big(T_k(x)\big) \big|\det \mathrm{J}_{T_k}(x)\big|}{\psi(x)} \psi(x) \mathrm{d}x \\
&= \sum_{k=1}^{K} \int_{\mathbb{R}^d} \frac{w_k\big(T_k(x)\big) q\big(T_k(x)\big) \big|\det \mathrm{J}_{T_k}(x)\big|}{\psi(x)} \mathbb{1}_{T_k^{-1}(B)}(x) \psi(x) \mathrm{d}x \\
&= \int_{\mathbb{R}^d} \underbrace{\sum_{k=1}^{K} \frac{w_k\big(T_k(x)\big) q\big(T_k(x)\big) \big|\det \mathrm{J}_{T_k}(x)\big|}{\psi(x)} \mathbb{1}_{T_k^{-1}(B)}(x)}_{\overrightarrow{\Pi}(x,B)} \psi(x) \mathrm{d}x,
\end{aligned}$$

so $\overrightarrow{\Pi}(x, B) = \sum_{k=1}^{K} v_k(x) \mathbb{1}\big(T_k(x) \in B\big)$ where $v_k(x)$ is given by:

$$v_k(x) = \frac{w_k\big(T_k(x)\big) q\big(T_k(x)\big) \big|\det \mathrm{J}_{T_k}(x)\big|}{\sum_{j=1}^{K} w_j\big(T_j(x)\big) q\big(T_j(x)\big) \Big|\det \mathrm{J}_{T_j}(x)\Big|}. \tag{A.2}$$

A.5 Derivation of GEM objective

In this section we will explicit model parameters at step t using superscript as in $\psi^{(\theta_t)}(x)$. First, for purpose of conciseness, let us define a subsidiary function which describes the joint pdf of $(\widetilde{X}, U)$ for model parameters θ:

$$h_k^{(\theta)}(x) = w_k^{(\theta)}\Big(T_k^{(\theta)}(x)\Big) q\Big(T_k^{(\theta)}(x)\Big) |\det J_{T_k^{(\theta)}}(x)|, \forall x \in \mathbb{R}^d \text{ and } k = 1, ..., K.$$

Since $\log\big(\psi^{(\theta)}(x)\big) = \mathbb{E}_\rho\Big[\log\big(\psi^{(\theta)}(x)\big)\Big]$, where the expectation is taken with respect to discrete categorical rv U with a probability measure ρ such that $U \sim \text{Categorical}\big(\rho(1), ..., \rho(K)\big)$, we can write:

$$\begin{aligned}
\log\big(\psi^{(\theta)}(x)\big) &= \mathbb{E}_\rho\Big[\log\Big(h_U^{(\theta)}(x)\Big) - \log\Big(\Pr^{(\theta)}\big(U|x\big)\Big)\Big] \\
&= \mathbb{E}_\rho\Big[\log\Big(h_U^{(\theta)}(x)\Big) - \log\big(\rho(U)\big)\Big] \\
&\quad + \mathbb{E}_\rho\Big[\log\big(\rho(U)\big) - \log\Big(\Pr^{(\theta)}\big(U|x\big)\Big)\Big].
\end{aligned}$$

The last term is $D_{\mathrm{KL}}\Big(\rho(U)||\Pr^{(\theta)}\big(U|x\big)\Big) \geq 0$, hence we have:

$$\log\big(\psi^{(\theta)}(x)\big) \geq \mathbb{E}_\rho\Big[\log\Big(h_U^{(\theta)}(x)\Big) - \log\big(\rho(U)\big)\Big], \tag{A.3}$$

where equality holds if and only if

$$\rho(k) = \Pr^{(\theta)}\big(U = k|x\big) \text{ for all values } k = 1, ..., K. \tag{A.4}$$

Let us finally turn to an iterative optimization scheme. Let θ_t be the current parameter. Let us set $\rho(k) = \Pr^{(\theta_t)}\big(U = k|x\big) = \Pr^{(\theta_t)}\big(z = T_k(x)|x\big) \overset{(A.2)}{=} v_k^{(\theta_t)}(x)$ for all $k = 1, ..., K$, and let us sum for $x = x_1, ..., x_M$. The rhs of (A.3) yields a function $g_{\theta_t}(\theta)$ which reads:

$$g_{\theta_t}(\theta) = \sum_{i=1}^{M}\sum_{k=1}^{K} v_k^{(\theta_t)}(x_i) \log\left(\frac{h_k^{(\theta)}(x_i)}{v_k^{(\theta_t)}(x_i)}\right),$$

and satisfies

$$g_{\theta_t}(\theta) \overset{(A.3)}{\leq} \sum_{i=1}^{M} \log\big(\psi^{(\theta)}(x)\big) \text{ for all } \theta \in \Theta, \tag{A.5}$$

$$g_{\theta_t}(\theta_t) \overset{(A.4)}{=} \sum_{i=1}^{M} \log\big(\psi^{(\theta_t)}(x_i)\big). \tag{A.6}$$

Therefore, if we compute θ_{t+1} via a GA step, (or, more generally, any method which ensures that $g_{\theta_t}\big(\theta_{t+1}\big) \geq g_{\theta_t}(\theta_t)\big)$, then by construction, we increase the log-likelihood of data $\big\{x_1, ..., x_M\big\}$ under ψ since:

$$\sum_{i=1}^{M} \log\big(\psi^{(\theta_{t+1})}(x_i)\big) \overset{(A.5)}{\geq} g_{\theta_t}\big(\theta_{t+1}\big) \overset{GA}{\geq} g_{\theta_t}(\theta_t) \overset{(A.6)}{=} \sum_{i=1}^{M} \log\big(\psi^{(\theta_t)}(x_i)\big).$$

Finally in our case, we can indeed check that the gradient of the surrogate function coincides with that of the target distribution:

$$\begin{aligned}
\nabla_\theta g_{\theta_t}(\theta)|_{\theta=\theta_t} &= \sum_{i=1}^{M}\sum_{k=1}^{K} v_k^{(\theta_t)}(x_i) \nabla_\theta \log\Big(h_k^{(\theta)}(x_i)\Big)|_{\theta=\theta_t} \\
&= \sum_{i=1}^{M}\sum_{k=1}^{K} v_k^{(\theta_t)}(x_i) \frac{\nabla_\theta h_k^{(\theta)}(x_i)}{h_k^{(\theta)}(x_i)}|_{\theta=\theta_t} = \sum_{i=1}^{M}\sum_{k=1}^{K} v_k^{(\theta_t)}(x_i) \frac{\nabla_\theta h_k^{(\theta)}(x_i)|_{\theta=\theta_t}}{h_k^{(\theta_t)}(x_i)} \\
&= \sum_{i=1}^{M} \frac{1}{\psi^{(\theta_t)}(x_i)} \sum_{k=1}^{K} \nabla_\theta h_k^{(\theta)}(x_i)|_{\theta=\theta_t} = \sum_{i=1}^{M} \frac{1}{\psi^{(\theta_t)}(x_i)} \nabla_\theta \sum_{k=1}^{K} h_k^{(\theta)}(x_i)|_{\theta=\theta_t} \\
&= \sum_{i=1}^{M} \frac{\nabla_\theta \psi^{(\theta)}(x_i)|_{\theta=\theta_t}}{\psi^{(\theta_t)}(x_i)} = \sum_{i=1}^{M} \nabla_\theta \log\big(\psi^{(\theta)}(x_i)\big)|_{\theta=\theta_t} = \nabla_\theta \sum_{i=1}^{M} \log\big(\psi^{(\theta)}(x_i)\big)|_{\theta=\theta_t},
\end{aligned}$$

which validates our construction of functions $\big\{g_{\theta_t}\big\}_{t=1,2,...}$

A.6 Cascading DIF in practice

We now see that the cascaded models discussed in section 4.3.10 can be implemented efficiently for both the VI and VDE problems. We explicit here the according objectives to be optimized for a cascade of two DIF; but with using recursion, this construction can of course be extended to more that two DIF.

VI

First, we can obtain samples $\widetilde{X}$ from Ψ by sequentially applying $Z_1 \sim \overleftarrow{\Pi}^{[1]}(Z)$ and then $\widetilde{X} \sim \overleftarrow{\Pi}^{[1]}(Z_1)$ to original samples $Z \sim Q$. This corresponds to the following sampling scheme:

$$\widetilde{X} = T_{U_0}^{[0]^{-1}}\Big(T_{U_1}^{[1]^{-1}}(Z)\Big) \text{ where } Z \sim Q, U_1 \sim \text{Categorical}\Big\{w_{k_1}^{[1]}(Z)\Big\}_{k_1=1,\dots,K_1}$$
$$\text{and } U_0 \sim \text{Categorical}\Big\{w_{k_0}^{[0]}\Big(T_{U_1}^{[1]^{-1}}(Z)\Big)\Big\}_{k_0=1,\dots,K_0}.$$

We can use RB in a sequential manner in order to build a differentiable MC approximation of the reverse D_{KL}:

$$D_{\text{KL}}\big(\Psi||P\big) \approx \frac{1}{M}\sum_{\substack{i=1\\ z_i\sim Q}}^{M}\underbrace{\sum_{k_1=1}^{K_1} w_{k_1}^{[1]}(z_i)\underbrace{\sum_{k_0=1}^{K_0} w_{k_0}^{[0]}(T_{k_1}^{[1]^{-1}}(z_i))\log\left(\frac{\psi(T_{k_0}^{[0]^{-1}}(T_{k_1}^{[1]^{-1}}(z_i)))}{p(T_{k_0}^{[0]^{-1}}(T_{k_1}^{[1]^{-1}}(z_i)))}\right)}_{\mathbb{E}\left[J|Z=z_i,U_1\right]}}_{\mathbb{E}\left[J|Z=z_i\right]} \tag{A.7}$$

which, as explained in section 4.3.6, corresponds to an RB approximation where we successively marginalized out the Categorical latent variables U_0 and U_1.

VDE

Next, the pdf induced by this cascade model can be easily computed via the following recursion:

$$\psi = \mathcal{F}\left(\mathcal{F}\left(q; \overleftarrow{\Pi^{[1]}}\right); \overleftarrow{\Pi^{[0]}}\right),$$

hence, as explained in section 4.3.7, one can use this model for VDE by maximizing the log-likelihood. Alternately, one can maximize a GEM surrogate, which reads

$$g_{\theta_t}(\theta) = \sum_{\substack{i=1\\ x_i\sim P}}^{M}\sum_{k_0=1}^{K_0} v_{k_0}^{[0]^{(\theta_t)}}(x_i)\sum_{k_1=1}^{K_1} v_{k_1}^{[1]^{(\theta_t)}}(T_{k_0}^{[0]^{(\theta_t)}}(x_i))\log\left(\frac{h_{k_0,k_1}^{(\theta)}(x_i)}{v_{k_0}^{[0]^{(\theta_t)}}(x_i)v_{k_1}^{[1]^{(\theta_t)}}(T_{k_0}^{[0]^{(\theta_t)}}(x_i))}\right)$$

where

$$h_{k_0,k_1}(x_i) = w_{k_0}^{[0]}(T_{k_0}^{[0]}(x_i))w_{k_1}^{[1]}(T_{k_1}^{[1]}(T_{k_0}^{[0]}(x_i)))$$
$$\times\, q\Big(T_{k_1}^{[1]}\Big(T_{k_0}^{[0]}(x_i)\Big)\Big)|\det \mathrm{J}_{T_{k_0}^{[0]}}(x_i)|.|\det \mathrm{J}_{T_{k_1}^{[1]}}\Big(T_{k_0}^{[0]}(x_i)\Big)|.$$

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Contributions à l'apprentissage statistique de lois a posteriori pour l'inférence bayésienne sans vraisemblance

Mots clés : Apprentissage Automatique, Inférence Bayésienne, Likelihood-free, Echantillonnage et estimation Monte Carlo, Quantification d'incertitude

Résumé : L'inférence bayésienne a posteriori est utilisée dans de nombreuses applications scientifiques et constitue une méthodologie répandue pour la prise de décision en situation d'incertitude. Elle permet aux praticiens de confronter les observations du monde réel à des modèles d'observation pertinents, et, en retour, d'inférer la distribution d'une variable explicative. Dans de nombreux domaines et applications pratiques, nous considérons des modèles d'observation de plus en plus complexes pour leur pertinence scientifique, mais au prix de densités de probabilité incalculables. En conséquence, à la fois la vraisemblance et la distribution a posteriori sont indisponibles, rendant l'inférence a posteriori à l'aide des méthodes de Monte Carlo habituelles irréalisable.

Dans ce travail nous supposons que le modèle d'observation nous génère un jeu de données, et le contexte de cette thèse est de coupler l'inférence Bayésienne à l'apprentissage statistique afin de pallier cette limitation et permettre l'inférence a posteriori dans le cadre likelihood-free. Ce problème, formulé comme l'apprentissage d'une distribution a posteriori, inclut les tâches habituelles de classification et de régression, mais il peut également être une alternative aux méthodes "Approximate Bayesian Computation" dans le contexte de l'inférence basée sur la simulation, où le modèle d'observation est plutôt un modèle de simulation avec une densité implicite.

L'objectif de cette thèse est de proposer des contributions méthodologiques pour l'apprentissage bayésien a posteriori. Plus précisément, notre objectif principal est de comparer différentes méthodes d'apprentissage dans le cadre de l'échantillonnage Monte Carlo et de la quantification d'incertitude.

Nous considérons d'abord l'approximation a posteriori basée sur le "likelihood-to-evidence-ratio", qui a l'avantage principal de transformer un problème d'apprentissage de densité conditionnelle en un problème de classification binaire. Dans le contexte de l'échantillonnage Monte Carlo, nous proposons une méthodologie pour échantillonner suivant la distribution résultante d'une telle approximation a posteriori. Pour résumer notre contribution : nous tirons parti de la structure sous-jacente du modèle, compatible avec les algorithmes d'échantillonnage usuels basés sur un quotient de densités, pour obtenir des procédures d'échantillonnage simples, sans hyperparamètre et ne nécessitant d'évaluer aucune fonction de densité.

Nous nous tournons ensuite vers le problème de la quantification de l'incertitude épistémique. D'une part, les modèles normalisés, tels que la construction discriminante, sont faciles à appliquer dans le contexte de la quantification de l'incertitude bayésienne. D'autre part, bien que les modèles non normalisés, comme le likelihood-to-evidence-ratio, ne soient pas facilement applicables dans les problèmes de quantification d'incertitude épistémique, une construction non normalisée spécifique, que nous appelons générative, est effectivement compatible avec la quantification de l'incertitude bayésienne via la distribution prédictive a posteriori. Dans ce contexte, nous expliquons comment réaliser cette quantification de l'incertitude dans les deux techniques de modélisation, générative et discriminante, puis nous proposons une comparaison des deux constructions dans le cadre de l'apprentissage bayésien.

Enfin nous abordons le problème de la modélisation paramétrique avec densité tractable, qui est effectivement une exigence pour la quantification de l'incertitude épistémique dans les méthodes de modélisations générative et discriminante. Nous proposons une nouvelle construction d'un modèle paramétrique, qui est une double extension des modèles de mélange et des flots normalisants. Ce modèle peut être appliqué à de nombreux types de problèmes statistiques, tels que l'inférence variationnelle, l'estimation de densité et de densité conditionnelle, car il bénéficie d'une évaluation rapide et exacte de la fonction de densité, d'un schéma d'échantillonnage simple, et d'une approche de reparamétrisation des gradients.

Title : Contributions to posterior learning for likelihood-free Bayesian inference

Keywords : Statistical learning, Bayesian Inference, Likelihood-free, Monte Carlo sampling and estimation, Uncertainty quantification

Abstract : Bayesian posterior inference is used in many scientific applications and is a prevalent methodology for decision-making under uncertainty. It enables practitioners to confront real-world observations with relevant observation models, and in turn, infer the distribution over an explanatory variable. In many fields and practical applications, we consider ever more intricate observation models for their otherwise scientific relevance, but at the cost of intractable probability density functions. As a result, both the likelihood and the posterior are unavailable, making posterior inference using the usual Monte Carlo methods unfeasible.

In this thesis, we suppose that the observation model provides a recorded dataset, and our aim is to bring together Bayesian inference and statistical learning methods to perform posterior inference in a likelihood-free setting. This problem, formulated as learning an approximation of a posterior distribution, includes the usual statistical learning tasks of regression and classification modeling, but it can also be an alternative to Approximate Bayesian Computation methods in the context of simulation-based inference, where the observation model is instead a simulation model with implicit density.

The aim of this thesis is to propose methodological contributions for Bayesian posterior learning. More precisely, our main goal is to compare different learning methods under the scope of Monte Carlo sampling and uncertainty quantification. We first consider the posterior approximation based on the likelihood-to-evidence ratio, which has the main advantage that it turns a problem of conditional density learning into a problem of binary classification. In the context of Monte Carlo sampling, we propose a methodology for sampling from such a posterior approximation. We leverage the structure of the underlying model, which is conveniently compatible with the usual ratio-based sampling algorithms, to obtain straightforward, parameter-free, and density-free sampling procedures.

We then turn to the problem of uncertainty quantification. On the one hand, normalized models such as the discriminative construction are easy to apply in the context of Bayesian uncertainty quantification. On the other hand, while unnormalized models, such as the likelihood-to-evidence-ratio, are not easily applied in uncertainty-aware learning tasks, a specific unnormalized construction, which we refer to as generative, is indeed compatible with Bayesian uncertainty quantification via the posterior predictive distribution. In this context, we explain how to carry out uncertainty quantification in both modeling techniques, and we then propose a comparison of the two constructions under the scope of Bayesian learning.

We finally turn to the problem of parametric modeling with tractable density, which is indeed a requirement for epistemic uncertainty quantification in generative and discriminative modeling methods. We propose a new construction of a parametric model, which is an extension of both mixture models and normalizing flows. This model can be applied to many different types of statistical problems, such as variational inference, density estimation, and conditional density estimation, as it benefits from rapid and exact density evaluation, a straightforward sampling scheme, and a gradient reparameterization approach.

Institut Polytechnique de Paris
91120 Palaiseau, France