Divergence minimization: sampling, approximate inference, and more

Thomas Guilmeau[1,a], Emilie Chouzenoux[1,b], and Víctor Elvira[2]

[1]Université Paris-Saclay, CentraleSupélec, INRIA, CVN, France
[a] thomas.guilmeau@inria.fr
[b] emilie.chouzenoux@centralesupelec.fr
[2]School of Mathematics, University of Edinburgh, United Kingdom
victor.elvira@ed.ac.uk

Abstract

Keywords. Variational inference, Adaptive importance sampling.

MSC2020 Subject Classification.

1 Introduction

Situations where a complicated π arises. Bayesian statistics and machine learning at large, signal processing, Bayesian neural network,...

Perform approximate inference on π, estimate integrals, moments, normalization constants,... Most of the times, this is done using an approximation $q \in \mathcal{Q}$ of π. We can measure the discrepancy between π and its approximation using a divergence D. The best approximation can then be found be solving an optimization problem of the form

$$\underset{q \in \mathcal{Q}}{\text{minimize}}\, D(\pi, q). \tag{$P_{D,\mathcal{Q}}$}$$

We will show that problems of the form of Problem ($P_{D,\mathcal{Q}}$) arise in many settings, including variational inference and importance sampling. Since problems of the form ($P_{D,\mathcal{Q}}$) are central, we then review many instances of it from different fields with algorithm to solve it.

Notation We will work in a measurable space $\mathcal{X}$ with its Borel algebra $\mathcal{B}(\mathcal{X})$. The set of measures over $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ will be denoted by $\mathcal{M}(\mathcal{X})$, and the set of probability measures over $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ by $\mathcal{P}(\mathcal{X})$. Given a measure $\nu \in \mathcal{M}(\mathcal{X})$, we denote the set of (probability) measures over $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ that are absolutely continuous with respect to ν by $\mathcal{M}(\mathcal{X}, \nu)$, and $\mathcal{P}(\mathcal{X}, \nu)$ in the case of probability measures. The Lebesgue measure will be denoted by dx.

Contents

2 Important situations where Problem ($P_{D,\mathcal{Q}}$) arises

We now review several tasks in statistics where practitioners are faced with instances of Problem ($P_{D,\mathcal{Q}}$). We start with situations which are explicit instances of ($P_{D,\mathcal{Q}}$), including variational inference, information projection, and importance sampling, before introducing more intricate tasks where problems of the form of ($P_{D,\mathcal{Q}}$) appear.

The tasks that we review below use the Kullback-Leibler divergence, denoted by KL, as well as the Rényi and α-divergences, denoted respectively by RD_α and D_α, with $\alpha > 0$. These divergences are introduced with more details in Section 3.1. For the moment, one need only to keep in mind that these divergences are non-negative and equal to zero if and only their two arguments are equal. Thus, they allow to get a sense of the discrepancy between two probability densities.

2.1 Variational inference

Variational inference is an approach to statistical inference, and especially Bayesian inference, that resolves around trying to find the best approximation to a target density, usually a Bayesian posterior, within a family of approximating distributions.

For the sake of illustration, consider the following Bayesian statistics example: suppose that we have N observations $\{y_n\}_{n=1}^N$ of the random variable Y that have the likelihood $y \longmapsto p(y|X)$ given the latent variable X, and that we have a prior p_0 on X. Then, we consider the Bayesian posterior on X, which has density

$$p(x|\{y_n\}_{n=1}^N) = \frac{p(\{y_n\}_{n=1}^N|x)p_0(x)}{\int p(\{y_n\}_{n=1}^N|x)p_0(x)m(dx)}, \forall x \in \mathcal{X}. \tag{1}$$

In most cases, the normalization constant of this distribution is intractable due to high dimension.

In order to infer information about $X|\{y_n\}_{n=1}^N$, variational inference methods aim at approximating the density $x \longmapsto p(x|\{y_n\}_{n=1}^N)$ by an approximation p chosen in some set $\mathcal{Q}$. The set $\mathcal{Q}$ of approximating distribution is often chosen such that its members are easy to deal with. Thus, mean-field models have long been prevalent in variational inference [16], but more expressive approximating families $\mathcal{Q}$ have been increasingly used in the recent years [106]. Most of the times, the best approximating distribution is searched as the minimizer of the exclusive KL divergence (see Section 3.1.1 and Section 3.2.1 for more details about this divergence and the choice of terminology) to the target $\pi = p(\cdot|\{y_n\}_{n=1}^N)$, leading to the divergence-minimization problem

$$\underset{p \in \mathcal{Q}}{\text{minimize}}\, KL(p, \pi).$$

In most applications of interest, π is only known up to a normalization constant, as it is the case for the Bayesian posterior (1). This means that we only have access to $\widetilde{\pi}$ which is such that $\pi = \frac{1}{Z_\pi}\widetilde{\pi}$ for some normalization constant $Z_\pi > 0$. In the case of the posterior (1), $Z_\pi = \int p(\{y_n\}_{n=1}^N|x)p_0(x)m(dx)$ and is often called the model evidence or the marginal likelihood and plays a central role in hypothesis and model testing [75]. Knowledge of Z_π is not needed to minimize $KL(\cdot, \pi)$, as one can decompose the KL divergence into separate terms, some involving p and some other involving Z_π as follows [16, 106].

$$\log Z_\pi = KL(p, \pi) + \int \log\left(\frac{\widetilde{\pi}(x)}{p(x)}\right) p(x)m(dx). \tag{2}$$

Therefore, minimizing the exclusive KL divergence is equivalent to maximizing the so-called evidence lower bound, which is defined as

$$p \longmapsto \int \log(\widetilde{\pi}(x))p(x)m(dx) - \int \log(p(x))p(x)m(dx). \tag{3}$$

This quantity can be unbiasedly approximated using samples from p and due to (2) and the non-negativity of the KL divergence, it lower-bounds $\log Z_\pi$.

Other divergences have also been used to approximate the distribution of interest π. We detail here the work of [72] which proposed to solve the divergence-minimization problem

$$\underset{p \in \mathcal{Q}}{\text{minimize}}\, RD_\alpha(p, \pi),$$

with $\alpha \in (0,1) \cup (1, +\infty)$ and RD_α being the Rényi divergence (see Section 3.1.2 for more details about this divergence). The resulting optimization problem also comes with a quantity, called the variational Rényi bound, that is decoupled from Z_π. Indeed, it was shown in [72] that we can write

$$\log Z_\pi = RD_\alpha(p, \pi) + \frac{1}{1-\alpha} \log \left(\int \left(\frac{\widetilde{\pi}(x)}{p(x)} \right)^{1-\alpha} p(x) m(dx) \right). \tag{4}$$

The variational Rényi bound is the function

$$p \longmapsto \frac{1}{1-\alpha} \log \left(\int \left(\frac{\widetilde{\pi}(x)}{p(x)} \right)^{1-\alpha} p(x) m(dx) \right), \tag{5}$$

and minimizing the Rényi divergence is equivalent to maximizing the variational Rényi bound. Similarly to the evidence lower bound, the variational Rényi bound is a lower bound of $\log Z_\pi$. However, approximating it with samples from the approximation p will lead to biased approximations due to the composition with the logarithm.

The choice of statistical divergence to minimize is subject to many considerations. First, for each divergence, the corresponding optimal approximation will have different properties. We review these properties more thoroughly in Section 3.2.1, but let us mention here a few results. The exclusive KL divergence produces mode-seeking approximations that under-estimate the variance of the posterior [77]. Depending on the parameter α the Rényi divergences favour different types of approximation, ranging from mode-fitting approximations that will concentrate around the mode of the target to mass-covering approximations that will cover all the mass of the target [72].

A second consideration that must be taken into account is the difficulty of reaching this optimal approximating distributions. In particular, using divergences other than the exclusive KL could create better solutions, but lead to algorithms that have bias and may make optimization difficult, especially as dimension grows [42, 31]. Finally, another aspect that comes into play in the big-data regime is the possibility to perform data-subsampling, which is natural when minimizing the exclusive KL divergence [52].

2.2 Information projection

Information projection is an operation similar to standard projection that consists in projecting a distribution π to a set $\mathcal{Q}$ by finding $q \in \mathcal{Q}$ that is the closest to π in terms of a given divergence D. Information projection is thus an instance of Problem ($P_{D,\mathcal{Q}}$).

Definition 1. The distribution $q \in \mathcal{Q}$ that solves

$$\underset{q \in \mathcal{Q}}{\text{minimize}}\, KL(q, \pi) \tag{6}$$

is the information projection of π onto the set $\mathcal{Q}$.

The information projection problem involves the exclusive KL divergence and has been studied for instance in [37, 28]. Solving this problem has a wide range of applications in hypothesis testing and large deviation theory, among other uses (see [28] and references therein).

The information projection problem is very close to the variational inference problem, and similarly, projection with other divergences have been considered as well. For instance, the reverse information projection, which consists in minimizing $q \longmapsto KL(\pi, q)$ over $\mathcal{Q}$.

Definition 2. The distribution $q \in \mathcal{Q}$ that solves

$$\underset{q\in\mathcal{Q}}{\text{minimize}}\, KL(q, \pi) \tag{7}$$

is the reverse information projection of π onto the set $\mathcal{Q}$.

The reverse information projection has links with maximum likelihood estimation [103, 28], as the maximum likelihood estimator tends to the reverse information projection when the number of available observations grows to infinity. The reverse information projection problem is also relevant in hypothesis testing [44].

Finally, note that the α- and reverse α- projections, which are obtained by minimizing respectively $q \longmapsto RD_\alpha(q, \pi)$ and $q \longmapsto RD_\alpha(\pi, q)$ are also of interest. They have been studied in [101, 62], with applications in information theory and robust statistics.

2.3 Importance sampling

Importance sampling is a Monte Carlo procedure that allows to approximate integrals of the form

$$I = \int h(x)\pi(x)m(dx) \tag{8}$$

when samples from $\pi \in \mathcal{P}(\mathcal{X}, m)$ are unavailable or when it is undesirable to sample from π. An example of the former situation is when π is a Bayesian posterior as in Equation (1) and one wants to estimate, for instance, its normalization constant ($h(x) = 1$), its mean ($h(x) = x$), or its variance ($h(x) = x^2$). An example of the latter situation is when facing rare-event estimation, that is the estimation of I with $h = \mathbb{1}_C$ where $C \subset \mathcal{X}$ is such that $I = \pi(X \in C)$ is very low. In such situation, the number of samples required from π to construct a Monte Carlo approximation of I with reasonable variance is prohibitive and importance sampling can be used to circumvent this issue [10].

In order to avoid sampling from π, importance sampling resorts to a sampling distribution $q \neq \pi$, called a proposal distribution. As we will explain, having $\pi \neq q$ is not only practical when we cannot sample from π, but it can yield better performance than the plain Monte Carlo estimator. When the density of π can be evaluated, we can construct from the proposal p the unnormalised importance sampling estimator $\widehat{I}_{\text{UIS}}$ of I as

$$\widehat{I}_{\text{UIS}} = \frac{1}{N}\sum_{n=1}^{N}\frac{\pi(x_n)}{q(x_n)}h(x_n),\ x_n \overset{\text{i.i.d.}}{\sim} p,\ \forall n \in [\![1, N]\!]. \tag{9}$$

A more complicated situation arises when the density of π can only be evaluated up to a normalization constant, that is we can only evaluate $\widetilde{\pi} \in \mathcal{M}(\mathcal{X}, m)$ such that $\pi = \frac{1}{Z_\pi}\widetilde{\pi}$ for some normalization constant $Z_\pi > 0$. This is the case for instance in Bayesian statistics where π is the posterior (1). In this situation, we can use the so-called self-normalized importance sampling estimator $\widehat{I}_{\text{SNIS}}$, that jointly estimates $Z_\pi = \int \widetilde{\pi}(x)m(dx)$ and $\int h(x)\widetilde{\pi}(x)m(dx)$. More precisely, $\widehat{I}_{\text{SNIS}}$ is the estimator defined by

$$\widehat{I}_{\text{SNIS}} = \sum_{n=1}^{N}\overline{w}_n h(x_n),\ x_n \overset{\text{i.i.d.}}{\sim} q,\ \overline{w}_n = \frac{1}{\sum_{m=1}^{N}\frac{\widetilde{\pi}(x_m)}{q(x_m)}}\frac{\widetilde{\pi}(x_n)}{q(x_n)},\ \forall n \in [\![1, N]\!]. \tag{10}$$

For each $n \in [\![1, N]\!]$, the ratios $w_n = \frac{\pi(x_n)}{q(x_n)}$ are called the unnormalised importance weights, while their normalised version $\overline{w}_n$ are called the normalised importance weights.

If plain Monte Carlo amounts to choosing $q = \pi$, importance sampling allows a supplementary degree of freedom as we can select the proposal q. This raises the question of the selection of an appropriate proposal. The need for algorithms to construct good proposals gave rise to the adaptive importance sampling procedures, which aim at iteratively constructing good proposal distributions. Such algorithms were reviewed in [18]. We now review some theoretical results that can guide the choice of a good proposal q.

We have from [88, Theorem 9.1] that the unnormalised importance sampling estimator is unbiased, $\mathbb{E}[\widehat{I}_{\mathrm{UIS}}] = I$, and if $\pi(x) > 0$ implies $q(x) > 0$, that its variance is

$$\mathbb{V}[\widehat{I}_{UIS}] = \frac{1}{N}\left(\int \left(\frac{h(x)\pi(x)}{q(x)}\right)^2 q(x)m(dx) - I^2\right). \tag{11}$$

Remark that the condition on π and q requires q to cover the mass of π. In the particular case where h takes non-negative values and $I > 0$, we can introduce the probability distribution $\pi_h \in \mathcal{P}(\mathcal{X}, m)$ whose density is given by

$$\pi_h(x) = \frac{1}{I}\pi(x)h(x),\ \forall x \in \mathcal{X}. \tag{12}$$

We can remark that when $q = \pi_h$, $\mathbb{V}[\widehat{I}_{UIS}] = 0$, meaning that π_h can be seen as the optimal proposal to construct the importance sampling estimator. This result shows to which extent importance sampling can improve upon plain Monte Carlo. In the case of rare-event estimation, $h = \mathbb{1}_C$, and π_h is equal to the target π conditioned on the event C. Interestingly, we can rewrite the variance under a form recalling the χ^2 divergence D_2 (see Section 3.1.2 for more details about this divergence),

$$\mathbb{V}[\widehat{I}_{UIS}] = \frac{I^2}{N}\left(\int \left(\frac{\pi_h(x)}{q(x)}\right)^2 q(x)m(dx) - 1\right), \tag{13}$$

and thus show that minimizing the variance of $\widehat{I}_{UIS}$ amounts to solving a divergence-minimization problem coming from minimizing $q \longmapsto D_2(\pi, \cdot)$.

Controlling the variance of the estimator $\widehat{I}_{UIS}$ is sometimes seen as too demanding, with some authors preferring the use of high-probability bounds (see the discussion in the introduction of [97] for instance). These considerations have led to other divergence-minimization formulations of the choice of importance sampling proposal. We have from [80, Theorem 3.1] that the divergence $RD_\alpha(\pi, q)$ for $\alpha \in (1, 2]$ controls a high-probability bound on the performance of $\widehat{I}_{\mathrm{UIS}}$. Other results in the importance sampling literature give guidelines on the choice of the number of samples N so that $\widehat{I}_{\mathrm{UIS}}$ gives good estimates with high probability. In this line, it is assumed in [22, 97] that the proposal q is fixed and a necessary sample N is computed in terms of the discrepancy between π and q in the case when $h \equiv 1$. In particular, it is shown that N needs to be of order $N \approx \exp KL(\pi, q)$ or of order $N \approx D_2(\pi, q)$. Therefore, having q that minimizes these divergences ensure that the required number of samples is as low as possible.

Results about the performance of the self-normalized estimator $\widehat{I}_{\mathrm{SNIS}}$ are scarcer in the literature. First, this estimator is biased as $\mathbb{E}[\widehat{I}_{\mathrm{SNIS}}] \neq I$ in general. However, its mean-square error can be controlled for bounded integrands h using the following result adapted from [1, Theorem 2.1]:

$$\mathbb{E}\left[(\widehat{I}_{\mathrm{SNIS}} - I)^2\right] \leq \frac{4\|h\|_\infty^2}{N} D_2(\pi, q). \tag{14}$$

This result connects the construction of proposals for self-normalized importance sampling with the resolution of divergence-minimization problems. Let us also mention the result of [1, Theorem 2.3] that goes beyond bounded integrands, and the results about the necessary sample-size to prevent failure provided in [22] and given in terms of the KL divergence between the target and the proposal.

Generally, one wants to use proposals that have heavier tails than the target [88, Example 9.1] and among such a family minimise a divergence that induces mass-covering approximations of the target. We can also remark that all the aforementioned results imply that good proposals q minimise divergences that induce mass-covering approximations (see Section 3.2.1 for more details on this topic). This is not surprising in light of [88, Example 9.1] or considering that if $q(x) = 0$ where $\pi(x) > 0$, then the variance $\mathbb{V}[\widehat{I}_{UIS}]$ is infinite.

2.4 Other problems building on divergence-minimization

We review now some statistical tasks where an instance of Problem ($P_{D,\mathcal{Q}}$) intervenes as a building block or where the considered optimization problem is composite with one term being a divergence as in Problem ($P_{D,\mathcal{Q}}$).

2.4.1 Generalized Bayes

Bayesian posteriors can be obtained as the solution of divergence-minimization problem, a fact known as the Gibbs principle [60]. More explicitly, consider the posterior from (1), with prior p_0 and likelihood p with data $\{y_n\}_{n=1}^N$. Then, the posterior satisfies

$$p(\,\cdot\,|\{y_n\}_{n=1}^N) = \operatorname*{arg\,min}_{q \in \mathcal{P}(\mathcal{X})} \left(\mathbb{E}_{X\sim q}\left[-\sum_{n=1}^{N} \ln p(X|y_n) \right] + KL(q, p_0) \right). \tag{15}$$

This results means that the Bayesian posterior balances maximizing the log-likelihood of the data while minimizing the exclusive KL divergence to the prior.

In [60], a more general way of performing Bayesian inference is proposed, based on the so-called Rule of Three, which share common points with our Problem ($P_{D,\mathcal{Q}}$). The Rule of Three is obtained by considering a family of approximating distributions $\mathcal{Q}$, a general sample loss ℓ, and a divergence D that will measure the discrepancy to the prior p_0. This leads to considering the non-standard posterior distribution q^* obtained by

$$q^* = \operatorname*{arg\,min}_{q \in \mathcal{Q}} \left(\mathbb{E}_{X\sim q}\left[\sum_{n=1}^{N} \ell(X, y_n) \right] + D(p_0, q) \right). \tag{16}$$

Finding this generalized posterior is equivalent to solving an optimization that looks like an instance Problem ($P_{D,\mathcal{Q}}$) except that a linear term is added to the divergence.

Although, choosing $\mathcal{Q}$, ℓ, and D differently from (15) leads to a sub-optimal q^* that is different from the posterior of interest $p(\cdot|\{y_n\}_{n=1}^N)$, the authors of [60] argue that their Rule of Three can in fact lead to better performance in practice. Indeed, the initial problem may be ill-posed and allowing for more freedom could help to counteract this [60, Section 2.4].

2.4.2 Sparse solutions

When building an approximation q of π, it an be desirable to enforce some specific properties on the resulting approximation. These additional requirements can be enforced by imposing constraints, and thus further restrain the search space $\mathcal{Q}$, or by adding an additional term to the objective cost that will promote proposals with the sought properties. We mention briefly two examples where the approximations are required to exhibit some form of sparsity.

A sparse variational inference problem is formulated in [19] to address the task of designing a Bayesian core-set, that is a subset of the available data. This allows to deal with a smaller dataset and thus eases further inference in the context of Bayesian statistics. In this context, the target π is the full Bayesian posterior from (1), $\mathcal{Q}$ is an exponential family that depends on the prior and likelihood, and the core-set θ^* is obtained by

$$\theta^* = \arg\min_{q_\theta \in \mathcal{Q}} KL(q_\theta, \pi) + \iota_C(\theta), \tag{17}$$

where $C = \{\theta \in \Theta,\ \|\theta\|_0 \leq M\}$ for some pre-specified threshold $M > 0$.

The graphical lasso [51] can also be understood as a regularized instance of Problem ($P_{D,\mathcal{Q}}$). Indeed, suppose that we observe the empirical covariance $\widehat{C} \in \mathcal{S}_+^d$, then the graphical lasso estimator P_* is the solution to

$$\operatorname*{minimize}_{P \in \mathcal{S}_+^d} \operatorname{tr}(\widehat{C}P) - \log\det P + \lambda\|P\|_1. \tag{18}$$

Denote by π any Gaussian distribution with covariance $\widehat{C}$ and by $\mathcal{G}_0^d$ the family of Gaussian distributions in $\mathbb{R}^d$ with mean 0 (any fixed vector would work). Consider the distribution $q_{\Sigma_*} \in \mathcal{G}_0^d$ that solves

$$\operatorname*{minimize}_{q_\Sigma \in \mathcal{G}_0^d} KL(\pi, q_\Sigma) + \lambda\|\Sigma^{-1}\|_1. \tag{19}$$

Then, by inspecting the expression of the KL divergence between two Gaussians, one can observe that $\Sigma_*^{-1} = P_*$, highlighting the link between the graphical lasso and Problem ($P_{D,\mathcal{Q}}$).

2.4.3 Evolution strategies

Black-box global optimization problems are challenging optimization problems where one aims at minimizing an objective function $U : \mathcal{X} \to \mathbb{R}$ which is possibly non-smooth and non-convex having only access to zero-th order information about the objective (i.e., the gradients and higher-order information are unavailable). A successful strategy to solve such problems is to iteratively construct probability densities in some parametric family $\mathcal{Q}$ that will concentrate around regions of the search space where the objective function reaches low values. This is the basis of various methods, such as estimation-of-distributions algorithms [67] or evolution strategies such as the CMA-ES algorithm [50, 49]. We will show how the construction of such densities relate to divergence-minimization problems of the form ($P_{D,\mathcal{Q}}$).

The most direct way to search for parametric distributions that concentrate around the minimizers of U is to find

$$q^* = \arg\min_{q \in \mathcal{Q}} \mathbb{E}_{X\sim q}[U(X)].$$

The resolution of this optimization problem is considered for instance in [2]. More recently, [82] considered the minimization of the entropy-regularized function $q \longmapsto \mathbb{E}_{X\sim q}[U(X)] + \lambda \int \log(q(x))q(x)m(dx)$, aiming at promoting more exploration of the space. This problem is an instance of Problem ($P_{D,\mathcal{Q}}$), since it is equivalent to minimizing $q \longmapsto KL(\pi_\lambda^U, q)$ with π_λ^U being the Boltzmann distribution with temperature λ

$$\pi_\lambda^U(x) \propto \exp\left(\frac{1}{\lambda}U(x)\right), \forall x \in \mathcal{X}.$$

The two optimization problems that we mention lead to algorithms that explicitly use the values of U, while most efficient evolution strategies only use rankings of solutions. Therefore, rank-based variational formulations of the construction of parametric distributions q have been proposed [3, 104, 87]. These formulations can be formulated as iterated optimization problems, where given a current parametric density $q_k \in \mathcal{Q}$, $k \in \mathbb{N}$, the next density q_{k+1} is constructed as

$$q_{k+1} = \arg\min_{q \in \mathcal{Q}} \mathbb{E}_{X\sim q}[W_{q_k}^U(X)],$$

where $W^U_{q_k}$ is a ranking-based function introduced in [3, 87]. It was recently shown in [45] that this procedure amounts to solve at every iteration $k \in \mathbb{N}$ an instance of Problem ($P_{D,\mathcal{Q}}$) of the form

$$q_{k+1} = \arg\min_{q \in \mathcal{Q}} KL(\pi^U_{q_k}, q)$$

where $\pi^U_{q_k}$ is a tilted version of q_k using the values of $W^U_{q_k}$.

2.4.4 Maximum likelihood for models with latent variables

Suppose that one aims at maximising the log-likelihood for a model with latent variables, that is, we aim at maximising

$$\log p_\theta(x) = \int p_\theta(x|z)p(z)m(dz).$$

This is a common task, whose most famous example may be Gaussian mixture modelling, where $p_\theta(\cdot|z)$ is Gaussian and Z follows the uniform distribution over $[\![1, J]\!]$, J being the number of components in the mixture [51]. This problem is in general intractable as the expectation with respect to the latent variable z is difficult to compute.

In order to circumvent this issue, let us introduce a distribution with density q over the latent variable. Then, we can compute

$$\begin{aligned}\log p_\theta(x) &= \log p_\theta(x, z) - \log p_\theta(z|x)\\ &= \int \log p_\theta(x, z)q(z)m(dz) - \log p_\theta(z|z)q(z)m(dz)\\ &= \int \log\left(\frac{p_\theta(x, z)}{q(z)}\right) q(z)m(dz) - KL(q, p_\theta(\cdot|x)).\end{aligned}$$

Motivated by these computations, let us introduce the mapping $F : \Theta \times \mathcal{P}(\mathcal{Z}, m) \to \mathbb{R}$, defined by

$$F(\theta, q) = \int \log p_\theta(x, z)q(z)m(dz) - \int \log q(z)q(z)m(dz). \tag{20}$$

We now draw several observations. First, we can recognise that the mapping F can be interpreted as an evidence lower bound, the main difference with the evidence lower bound we introduced earlier being the dependence on the parameter θ. Second, if $KL(q, p_\theta(\cdot|x))$ is low, the evidence lower bound F will be closer to the true value of the log-likelihood. This shows the relationship between divergence-minimisation and maximum likelihood problems. Finally, our computations relates the maximisation of the evidence lower bound F with respect to its two variables, which is at the core of many methods that we will present in a moment, with the initial problem of maximising $\theta \longmapsto \log p_\theta(x)$.

One interpretation of expectation-maximization algorithms is that the perform alternating maximisations of the evidence lower bound [51]. At iteration $k \in \mathbb{N}$, one has (θ_k, q_k) and updates them through

$$\begin{aligned}q_{k+1} &= \arg\max_q \mathcal{L}(\theta_k, q),\\ \theta_{k+1} &= \arg\max_\theta \mathcal{L}(\theta, q_{k+1}).\end{aligned}$$

Remark that the first step can be rewritten as $q_{k+1} = \arg\min_q KL(q, p_{\theta_k}(\cdot|x))$, making the connection with divergence-minimisation explicit. It is obvious that at every iteration, $q_{k+1} = p_{\theta_k}(\cdot|x)$, and when this distribution is tractable, we recover usual expectation-maximisation updates.

In the situations where $q_{k+1} = p_{\theta_k}(\cdot|x)$ is not a tractable distribution, one can search for good distributions q by taking $q = q_\phi(\cdot|x)$. This is a form of amortization, since we are learning a mapping between the data and the latent variable, and this forms the core of variational autoencoders. A usual choice is to take

$$q_\phi(z|x) = \mathcal{N}(z; \mu_\phi(x), \sigma^2_\phi(x)),$$

where the mappings $x \longmapsto \mu_\phi(x), \sigma^2_\phi(x)$ are learnt be neural networks [15]. Then, one aims at maximising $(\theta, \phi) \longmapsto \mathcal{L}(\theta, q_\phi(\cdot|x))$.

Another approach to tackle situations where $q_{k+1} = p_{\theta_k}(\cdot|x)$ is intractable is to iteratively optimise F over $\Theta \times \mathcal{P}(\mathcal{X})$ where the sequence $\{q_k\}_{k\in\mathbb{N}}$ is only known through its samples. We detail this approach consisting in working with samples rather than their distributions, which falls under the *sampling as optimisation* umbrella, in Section 4.2. Such methods have been proposed for instance in [64, 5].

3 Statistical divergences

The optimization problem ($P_{D,\mathcal{Q}}$) arises from choosing a divergence D, that is a mapping $D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}_+$ such that $D(p_1, p_2) = 0$ if and only if $p_1 = p_2$, and an approximating family, that is a subset $\mathcal{Q} \subset \mathcal{P}(\mathcal{X}, \nu)$. The target π is often dictated by the problem at hand. For some problems, there are natural choices for D and $\mathcal{Q}$, but sometimes, the practitioners have to make a choice about which divergence and approximating family they will consider. We thus discuss these choice in this section. and important examples where such problems arise.

3.1 Introducing statistical divergences

Statistical divergences are functions allowing to measure the discrepancy between two probability distributions. They now have an utmost importance in statistics and machine learning. We first define several statistical divergences D that will be used throughout this work. Statistical divergences are mapping $D : \mathcal{P}(\mathcal{X}, m) \times \mathcal{P}(\mathcal{X}, m) \to \mathbb{R} \cup \{+\infty\}$ such that $D(p_1, p_2) \geq 0$ for any $p_1, p_2 \in \mathcal{P}(\mathcal{X}, m)$, with equality if and only if $p_1 = p_2$. However, divergences are not symmetric nor do they satisfy the triangle inequality in general. Therefore, they give a sense of the discrepancy between two given densities, but they are not proper distances on $\mathcal{P}(\mathcal{X}, m)$.

In the following, we introduce particular statistical divergences that will be the main focus of our review. Namely, we present the Kullback-Leibler divergence as well as the α- and Rényi divergences.

3.1.1 The Kullback-Leibler divergence

Definition 3. The Kullback-Leibler divergence, denoted by $KL(\cdot, \cdot)$, is defined for any probability densities $p_1, p_2 \in \mathcal{P}(\mathcal{X}, m)$ by

$$KL(p_1, p_2) = \int \log\left(\frac{p_1(x)}{p_2(x)}\right) p_1(x) m(dx), \tag{21}$$

using that $\log(p_1(x))p_1(x) = 0$ if $p_1(x) = 0$ (see [93, Definition 7.1] for more details on these singular cases).

Remark 1. In the context of divergence-minimization problems where one tries to approximate $\pi \in \mathcal{P}(\mathcal{X}, m)$, we will say that one minimizes the exclusive KL divergence when $KL(\cdot, \pi)$ is being minimized, and that one minimizes the inclusive KL divergence when $KL(\pi, \cdot)$ is being minimized.

We now give some properties of the Kullback-Leibler divergence, focusing in particular on its relations with other distances over probability measures.

Proposition 1. *The Kullback-Leibler divergence satisfies the following properties.*

(i) The KL divergence is a statistical divergence.

(ii) The KL divergence satisfies the Pinsker inequality

$$\frac{1}{2}\|p_1 - p_2\|_{TV}^2 \leq KL(p_1, p_2),$$

(iii) If p_1 satisfies a log-Sobolev inequality with constant $C > 0$, then

$$W_2(p_1, p_2) \leq \sqrt{2CKL(p_2, p_1)},\ \forall p_2 \in \mathcal{P}(\mathcal{X}),$$

where W_2 is the Wasserstein 2-distance.

3.1.2 The α- and Rényi divergences

Definition 4. We introduce the α- and Rényi divergences, respectively denoted by $D_\alpha(\cdot,\cdot)$, and $RD_\alpha(\cdot,\cdot)$, for some scalar $\alpha \in (0,1)\cup(1,+\infty)$. They are defined for any probability densities $p_1, p_2 \in \mathcal{P}(\mathcal{X}, m)$ by

$$D_\alpha(p_1,p_2) = \frac{1}{\alpha(\alpha-1)}\left(\int p_1(x)^\alpha p_2(x)^{1-\alpha} m(dx) - 1\right), \tag{22}$$

$$RD_\alpha(p_1,p_2) = \frac{1}{\alpha-1}\log\left(\int p_1(x)^\alpha p_2(x)^{1-\alpha} m(dx)\right). \tag{23}$$

The α-divergence recovers the Hellinger divergence when $\alpha = 0.5$ (which also happens to be a distance) as well as the χ^2 divergence when $\alpha = 2$. The Kullback-Leibler, α-, and Rényi divergences are closely related as the α- and Rényi divergences can be understood as deformations of the usual KL divergence. For a given parameter α, one can also relate the corresponding Rényi and α- divergences through $RD_\alpha(p_1,p_2) = \frac{1}{\alpha-1}\log\left(\alpha(\alpha-1)D_\alpha(p_1,p_2)+1\right)$.

We now present some properties of the $\alpha-$ and Rényi divergences, focusing on their relations with other measures of discrepancy and on the role of the parameter α.

Proposition 2 ([26, 101])**.** *Consider any two probability densities $p_1, p_2 \in \mathcal{P}(\mathcal{X}, m)$, then we have the following properties.*

(i) The α-, and Rényi divergences are statistical divergences: $KL(p_1,p_2), D_\alpha(p_1,p_2), RD_\alpha(p_1,p_2) \geq 0$ with equality if and only if $p_1 = p_2$ m-a.e.

(ii) We have the limits

$$RD_\alpha(p_1,p_2) \xrightarrow[\alpha\to 1]{} KL(p_1,p_2), \tag{24}$$

$$D_\alpha(p_1,p_2) \xrightarrow[\alpha\to 1]{} KL(p_1,p_2), \tag{25}$$

$$D_\alpha(p_1,p_2) \xrightarrow[\alpha\to 0]{} KL(p_2,p_1). \tag{26}$$

(iii) The mapping $\alpha \longmapsto RD_\alpha(p_1,p_2)$, which can be extended from $(0,1)\cup(1,+\infty)$ to $\mathbb{R}_{++}$ by writing $RD_1(p_1,p_2) = KL(p_1,p_2)$, is non-decreasing.

(iv) For any $\alpha \in (0,1)$, we have the symmetry properties

$$D_\alpha(p_1,p_2) = D_{1-\alpha}(p_2,p_2), \tag{27}$$

$$RD_\alpha(p_1,p_2) = \frac{\alpha}{1-\alpha}RD_{1-\alpha}(p_2,p_1). \tag{28}$$

(v) When $\alpha \in (0,1)$, Rényi divergences satisfy a Pinsker-like inequality of the form

$$\frac{\alpha}{2}\|p_1-p_2\|_{TV}^2 \leq RD_\alpha(p_1,p_2),\ \forall p_1,p_2 \in \mathcal{P}(\mathcal{X},m).$$

We illustrate the behaviour of Rényi divergences in Fig. 1 in the case of one-dimensional Gaussian distributions for varying α. In particular, we can observe that as α grows, the unit balls depicted in Fig. 1 gets smaller, showcasing the result of Proposition 2 *(iii)*. We can also observe that the shape of the balls is altered close to the boundary of the space of the parameters, that is for σ^2 being close to zero. This reflects that the geometry induced by the considered statistical divergences is different from the Euclidean one,

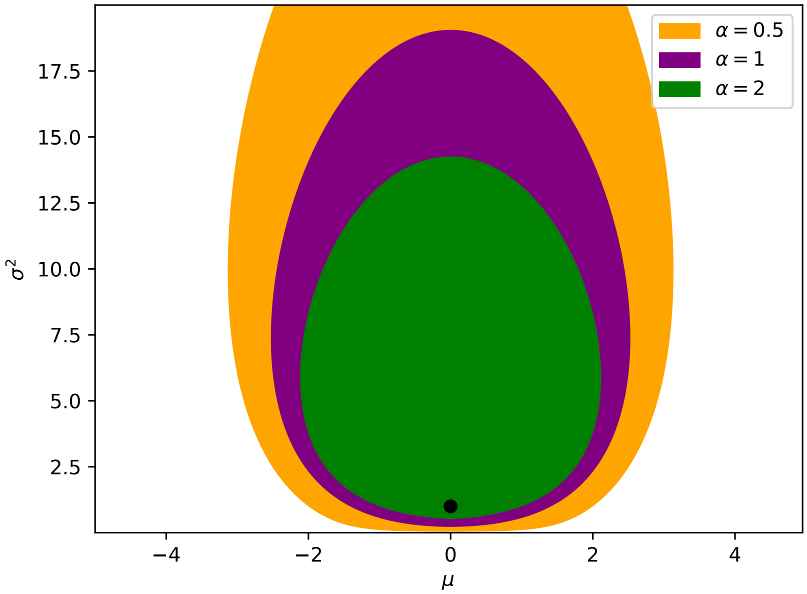


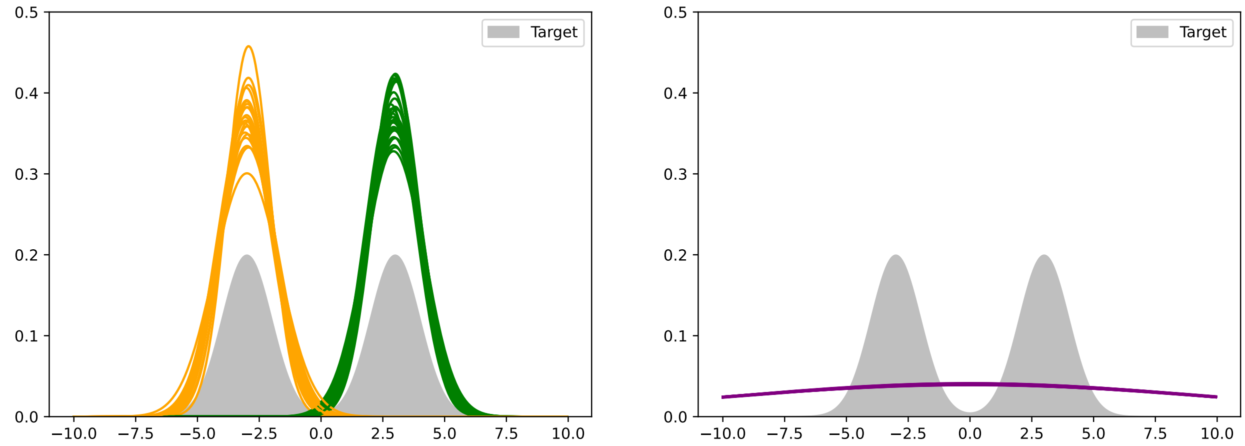
Figure 1: Unit balls in $\mathbb{R} \times \mathbb{R}_{++}$ of the form $\{(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{++},\ RD_\alpha(p_{0,1}, p_{\mu,\sigma^2}) \leq 1 \in\}$ for $\alpha \in \{0.5, 1, 2\}$, with $RD_1 = KL$. For $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{++}$, p_{μ,σ^2} is the one-dimensional Gaussian distribution with mean μ and variance σ^2 (see Section 4.1.1). The centre $(0, 1)$ of the balls, which corresponds to the reduced and centred Gaussian distribution, is marked by a black dot.

3.2 Properties of the minimisers induced by different divergences

The choice of the divergence D in the minimisation problem ($P_{D,\mathcal{Q}}$) impacts the properties of the optimal approximating distribution q^*. We review here some properties that one can expect from the optimal approximation, depending on the choice of the divergence.

3.2.1 Mass-covering and mode-fitting behaviours

In the case of multimodal targets, some divergences will promote a mode-fitting solution q^* while some other will promote a mass-covering behaviour in the solutions. In particular, minimising $KL(\cdot, \pi)$ has been known to lead to solutions that under-estimate the variance of π, as shown in [77] for mean-field Gaussian approximations. These properties have also been investigated in for a wider range of divergences in [70] in the context of GANs and in [76] in the context of mean-field Gaussian approximations. We illustrate these different behaviours in the context of the minimization of Rényi divergences between a target mixture and the Gaussian family in Figure 2.



(a) Here, $\alpha = 0.2$ and the problem is non-convex. We can observe two distinct stationary points, each corresponding to the approximation of one mode of the target.

(b) Here, $\alpha = 1$ and the problem is strictly convex. The only stationary point corresponds to a mass-covering approximating distribution.

Figure 2: Plots of the Gaussian approximations of a mixture-of-Gaussians target. The approximations are obtained by minimizing the Rényi divergence between the target and the approximating density. We show the final approximations obtained for 50 randomly-initialized runs of the algorithm from [46], where the approximation is randomly initialized at one of the component of the target mixture.

3.2.2 Pythagorean theorems

The information projection and α-projection have in common to benefit from a so-called Pythagorean theorems [28, 101] under convexity assumption on the set $\mathcal{Q}$. We reproduce here [101, Theorem 14] for α-projections (which generalizes the result for information projection) without the precise assumption. Suppose that $q_* \in \mathcal{Q}$ minimizes $q \longmapsto RD_\alpha(q, \pi)$, then for any $q \in \mathcal{Q}$

$$RD_\alpha(q, \pi) \geq RD_\alpha(\pi, q_*) + RD_\alpha(q_*, \pi). \tag{29}$$

This inequality can be verified to be analogous to the corresponding results in the Euclidean setting. Note also that the reverse information projection also benefits from similar results [28, Theorem 3.4].

4 Approximating families $\mathcal{Q}$

Facing Problem ($P_{D,\mathcal{Q}}$), one generally aims at choosing $\mathcal{Q}$ such that $\pi \in \mathcal{Q}$, in which case π is a solution of Problem ($P_{D,\mathcal{Q}}$), but this is often not possible in practice. We often expect the approximating distributions from $\mathcal{Q}$ to be tractable in the sense that it is easy to sample from them and to evaluate their density. These requirements, expressivity and tractability, are often competing.

We present in the following several classes of approximating families $\mathcal{Q}$. In Section 4.1, we first start with approximating families whose distributions can be described by finite-dimensional parameters, before moving in Section 4.2 to infinite-dimensional families.

4.1 Parametric families

We present in this section several families which can be described by finite-dimensional parameters, including Gaussian and Student distributions, exponential families, reparametrised families such as location-scale families and normalising flows, and mixture models.

4.1.1 Gaussian and Student distributions

We introduce in this section two well-known families of parametric distributions on $\mathbb{R}^d$, Gaussian and Student distributions. We review some of their properties as well as their connections.

Definition 5 (Gaussian family). The Gaussian family is the set $\mathcal{G}^d \subset \mathcal{P}(\mathbb{R}^d, dx)$ such that for each $q \in \mathcal{G}^d$, there exist unique $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathcal{S}^d_{++}$ satisfying

$$q(x) = \frac{1}{(2\pi)^{d/2}(\det \Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right), \forall x \in \mathbb{R}^d. \tag{30}$$

We will thus write $q = q_{\mu,\Sigma}$.

Gaussian distributions have well-defined moments, and are completely characterized by their first and second moments. For $q_{\mu,\Sigma} \in \mathcal{G}^d$, $q_{\mu,\Sigma}(X) = \mu$ and $q_{\mu,\Sigma}(XX^\top) = \Sigma + \mu\mu^\top$.

We now introduce the family of Student distributions, which are related to Gaussian distributions, but have heavier tails. This makes them useful in many situations, including statistical modelling [41, 92, 100, 21, 43, 68, 7] or importance sampling [20, 40, 102].

Definition 6 (Student family). The Student family is the set $\mathcal{T}^d \subset \mathcal{P}(\mathbb{R}^d, dx)$ such that for each $q \in \mathcal{T}^d$, there exist unique $\nu > 0$, $\mu \in \mathbb{R}^d$, and $\Sigma \in \mathcal{S}^d_{++}$ for which

$$q(x) = \frac{1}{Z_\nu} \det(\Sigma)^{-\frac{1}{2}} \left(1 + \frac{1}{\nu}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)^{-\frac{\nu+d}{2}} \tag{31}$$

with $Z_\nu = \frac{\Gamma(\nu/2)\nu^{d/2}\pi^{d/2}}{\Gamma((\nu+d)/2)}$, Γ denoting here the Gamma function We will thus write $q = q_{\mu,\Sigma,\nu}$ to denote the dependence of q on its parameters.

For a given $\nu > 0$, we also introduce the set $\mathcal{T}^d_\nu \in \mathcal{P}(\mathbb{R}^d, dx)$, which is the family of Student distributions with ν degrees of freedom. It is such that for any $q \in \mathcal{T}^d_\nu$, there exist $\mu \in \mathbb{R}^d$, and $\Sigma \in \mathcal{S}^d_{++}$ such that $q = q_{\mu,\Sigma,\nu}$.

Student distributions do not form an exponential family and tend to have heavy tails. Indeed, their first moment is defined only for $\nu > 1$, in which case we have $q_{\mu,\Sigma,\nu}(X) = \mu$ and their second moment only from $\nu > 2$, in which case we have $q_{\mu,\Sigma,\nu}(XX^\top) = \frac{\nu}{\nu-2}\Sigma + \mu\mu^\top$. When $\nu = 1$, we recover the family of Cauchy distributions, while Gaussian distributions are recovered in the light-tailed limit $\nu \to +\infty$. Indeed, consider

$\mu \in \mathbb{R}^d$ and $\Sigma \in \mathcal{S}_{++}^d$, with the associated Gaussian distribution $q_{\mu,\Sigma}$ and the Student distribution with $\nu > 0$ degrees of freedom $q_{\mu,\Sigma,\nu}$. Then, for any $x \in \mathbb{R}^d$, we have

$$q_{\mu,\Sigma,\nu}(x) \xrightarrow[\nu \to +\infty]{} q_{\mu,\Sigma}(x). \tag{32}$$

Another useful connection between Student and Gaussian distributions is that the former can be represented using the latter through a latent variable. Suppose that we have a random variable Z following a Gamma distribution with parameters $(\frac{\nu}{2}, \frac{\nu}{2})$. Then the random variable X such that $X|Z$ follows a Gaussian distribution with mean μ and covariance $\frac{1}{Z}\Sigma$ follows a Student distribution with ν degrees of freedom, mean μ, and scale Σ. Equivalently, for $\nu > 0$, $\mu \in \mathbb{R}^d$, and $\Sigma \in \mathcal{S}_{++}^d$,

$$q_{\mu,\Sigma,\nu}(x) = \int_0^{+\infty} q_{\mu,\frac{1}{z}\Sigma}(x) q_{\frac{\nu}{2},\frac{\nu}{2}}(z) dz, \, \forall x \in \mathbb{R}^d, \tag{33}$$

where for $z > 0$, $q_{\mu,\frac{1}{z}\Sigma}$ is the density of the Gaussian with mean μ and covariance $\frac{1}{z}\Sigma$ and $q_{\frac{\nu}{2},\frac{\nu}{2}}$ is the density of the Gamma distributions with parameters $(\frac{\nu}{2}, \frac{\nu}{2})$.

4.1.2 Exponential family

Exponential families are an important class of parametric probability distributions. They include many well-known discrete and continuous distributions such as Gaussian, Dirichlet, Wishart, Poisson, or Bernoulli distributions and benefit from a number of interesting properties that have been well-studied [17, 12]. These properties make exponential families useful models to work with, and they have been used in many tasks [11, 52, 56].

Definition 7 (Exponential family). $\mathcal{Q} \subset \mathcal{P}(\mathcal{X}, m)$ is an exponential family if there exists a sufficient statistics $t : \mathcal{X} \to \mathcal{H}$ for some finite dimensional Hilbert space $\mathcal{H}$ such that for each $q \in \mathcal{Q}$, there exists $\theta \in \mathcal{H}$ satisfying

$$q(x) = \exp\left(\langle \theta, t(x) \rangle - A(\theta)\right), \, \forall x \in \mathcal{X}, \tag{34}$$

where $A(\theta) := \log\left(\int \exp(\langle \theta, t(x) \rangle) m(dx)\right)$ is finite. If for any $q \in \mathcal{Q}$, there exists only a unique θ satisfying the above, we say that $\mathcal{Q}$ is a minimal exponential family, and we then write $q = q_\theta$.

We now recall some properties of exponential families that we will use throughout this thesis. These properties have a convex analysis flavour, and indeed, convex analysis provides a useful toolkit to study exponential families. We first recall results showing that the log-partition function can be shown to be a Legendre function. Legendre functions are an important class of functions in convex analysis, with many nice properties that we recall now.

Definition 8 (Legendre functions). A Legendre function is a function $h \in \Gamma_0(\mathcal{H})$ that is strictly convex on the interior of its domain $\operatorname{int}\operatorname{dom} h$, and essentially smooth. h is essentially smooth if it is differentiable on $\operatorname{int}\operatorname{dom} h$ and such that $||\nabla h(\theta_k)|| \xrightarrow[k \to +\infty]{} +\infty$ for every sequence $\{\theta_k\}_{k \in \mathbb{N}}$ converging to a boundary point of $\operatorname{dom} h$ with $\theta_k \in \operatorname{int}\operatorname{dom} h$ for every $k \in \mathbb{N}$.

Proposition 3 (Section 2.3 in [98]). *Let h be a Legendre function. Then we have that*

(i) *∇h is a bijection from $\operatorname{int}\operatorname{dom} h$ to $\operatorname{int}\operatorname{dom} h^*$, and $(\nabla h)^{-1} = \nabla h^*$.*

(ii) *The subdifferential ∂h of h is such that $\partial h(\theta) = \{\nabla h(\theta)\}$, $\forall \theta \in \operatorname{int}\operatorname{dom} h$ and $\partial h(\theta) = \emptyset$ if $\theta \notin \operatorname{int}\operatorname{dom} h$.*

(iii) *h is a Legendre function if and only if h^* is a Legendre function.*

(iv) *for every $\theta \in \operatorname{dom} h$, $\theta' \in \operatorname{int}\operatorname{dom} h$, the Bregman divergence $d_h(\theta, \theta') := h(\theta) - h(\theta') - \langle \nabla h(\theta), \theta' - \theta \rangle$ is non-negative and null if and only if $\theta = \theta'$.*

Proposition 4 ([17, 86, 12]). *Under the hypothesis that* $\operatorname{int}\operatorname{dom} A \neq \emptyset$*, the log-partition* A *is proper, lower semicontinuous, and strictly convex. In addition, all the partial derivatives of* A *exist on* $\operatorname{int}\operatorname{dom} A$ *and its gradient reads*

$$\nabla A(\theta) = q_\theta(t(X)),\ \forall \theta \in \operatorname{int}\operatorname{dom} A. \tag{35}$$

If $\mathcal{Q}$ *is minimal and steep (see [12, Chapter 8] for more details on these notions), then the log-partition function is a Legendre function. Under these assumptions, we have for any* $\theta \in \operatorname{int}\operatorname{dom} A$ *and* $\theta' \in \operatorname{dom} A$ *that*

$$KL(q_\theta, q_{\theta'}) = d_A(\theta', \theta).$$

Minimality ensures that each distribution in $\mathcal{Q}$ can be parametrized only by a unique parameter θ. Steepness is satisfied by most exponential families. It is in particular implied by having $\operatorname{dom} A$ being open [12, Theorem 8.2]. More precisely, the results of Proposition 4 come from [17, Theorem 1.13] for the convexity results, [12, Theorem 8.1] for the differentiability result, [12, Eq. (20)] for the steepness part, and [86] for the equivalence between the Bregman divergence and the KL divergence.

In the case of steep exponential families, the Legendre property on A allows to benefit from the results of Proposition 3, implying in particular that there is a bijection between the parameters θ and $\nabla A(\theta)$. Since $\nabla A(\theta) = q_\theta(t(X))$, the Legendre property on A leads to an alternative characterization of distributions of $\mathcal{Q}$ in terms of their sufficient statistics. This sheds another light on the Pitman-Koopman-Darbois theorem [99] which states that exponential families are the only families of distributions which have a sufficient statistics. Note that the quantities $q_\theta(t(X))$ are often called the moments of q_θ or the dual parameters of q_θ. Perhaps more surprisingly, the Bregman divergence induced by the Legendre function A admits a statistical interpretation that has been well-studied in the information geometry community [86]. Indeed, the Kullback-Leibler divergence between two distributions from $\mathcal{Q}$ is equivalent to the Bregman divergence d_A between their parameters, as we recalled in the above proposition.

We detail here the example of the Gaussian family, which forms an exponential family.

Example 1. Let $d \geq 1$ and consider the family of Gaussian distributions $\mathcal{G}^d$. Then $\mathcal{G}^d$ forms an exponential family [12], with sufficient statistics

$$t(x) = (x, xx^\top),\ \forall x \in \mathbb{R}^d. \tag{36}$$

Its corresponding parameters are $\theta = (\theta_1, \theta_2) \in \mathcal{H} = \mathbb{R}^d \times \mathcal{S}^d$ with $\theta_1 = \Sigma^{-1}\mu$, and $\theta_2 = -\frac{1}{2}\Sigma^{-1}$, showing that the natural parameters of an exponential family need not to be the ones commonly used. Its log-partition function is $A(\theta) = \frac{d}{2}\log(2\pi) - \frac{1}{4}\theta_1^\top \theta_2^{-1}\theta_1 - \frac{1}{2}\log\det(-2\theta_2)$. The domain of A is $\operatorname{dom} A = \mathbb{R}^d \times \left(-\mathcal{S}^d_{++}\right)$. The scalar product of $\mathcal{H}$ is taken as the sum of the scalar product of $\mathbb{R}^d$ and the one of $\mathcal{S}^d$.

4.1.3 Reparametrised families

We introduce now reparametrised families, which are parametric families of distributions for which the sampling can be decomposed into sampling from a base measure that is common to all distributions of the family, and then transforming the sample using a function that depend on the parameters of the considered distribution from the family.

Definition 9 (Reparametrised family). The parametric family $\mathcal{Q} = \{q_\theta \,|\, \theta \in \Theta\}$ is reparametrised if there exists a base measure q_0 such that for any $\theta \in \Theta$, there exists T_θ such that if $x \sim q_0$, $T_\theta(x) \sim q_\theta$.

Reparametrised family are central to modern black-box variational inference methods, as they allow to use the widely used *reparametrisation trick*. We discuss particular cases of reparametrised families.

Definition 10 (Location-scale family). If $\theta = (m, C)$ with $T_\theta(x) = Cx + m$, then $\mathcal{Q}$ is called a location-scale reparametrised family.

Location-scale reparametrised families include many well-known families such as the class of elliptical distributions, which includes in particular Gaussian and Student distributions.

We now introduce normalising flows, which are a particular case of reparametrised families introduced recently. Normalising flows are not location-scale reparemetrised families.

Definition 11 (Normalising flows). If T_θ is a diffeomorphism, then we call T_θ a normalising flow, and with a slight abuse of words, we will say that q_θ is a normalising flow and that $\mathcal{Q}$ a family of normalising flows.

Normalising flows allow for sampling, by sampling $x \sim q_0$ and computing $T_\theta(x) \sim q_\theta$, and they also allow for density evaluation [90]. Indeed, one can check that

$$q_\theta(x) = q_0(z)|\det \mathrm{Jac}_{T_\theta}(z)|^{-1} \text{ where } z = T_\theta^{-1}(x).$$

It is often more convenient, especially for implementation and optimisation, to express T_θ as a composition of several simple diffeomorphisms, that is, $\theta = (\theta_1, \dots, \theta_T)$ and $T_\theta = T_{\theta_T} \circ \dots \circ T_{\theta_1}$. This implies in particular that if $x \sim q_0$, then

$$(T_{\theta_T} \circ \dots \circ T_{\theta_1})(x) \sim q_\theta. \tag{37}$$

If each operator T_{θ_t} is a diffeomorphism, then their composition is also a diffeomorphism and it remains possible to compute the density $q_\theta(x)$ for any $x \in \mathcal{X}$.

In theory, normalising flows are able to approximate any target as soon as it is possible to construct a diffeomorphism between the chosen base measure and the target [90]. However, it is necessary to restrict normalising flows to particular architectures to have an efficient training process. This prevents normalizing flows to achieve the expressivity of theoretical models based on diffeomorphisms, as shown by recent results for normalising flows with realistic architectures. For instance, normalising flows with affine couplings are able to approximate any log-concave target [69]. [54] showed that many popular flow models are restricted to have the same tail properties as their base measures.

4.1.4 Mixture families

We first introduce mixture families, which are built from a given family $\mathcal{Q}$ and allow to create from there complex and expressive distributions while retaining some of the features of the original family $\mathcal{Q}$.

Definition 12 (Mixture families). Consider a family of probability distributions $\mathcal{Q} \in \mathcal{P}(\mathcal{X})$. We introduce the set of mixture of distributions from $\mathcal{Q}$ as

$$\mathrm{mixt}\mathcal{Q} = \left\{ \sum_{j=1}^{J} \lambda_j q_j \,|\, J \in \mathbb{N},\, \lambda \in \Delta^J,\, q_j \in \mathcal{Q},\, \forall j \in \{1, \dots, J\} \right\}, \tag{38}$$

and the set of J-mixtures of distributions from $\mathcal{Q}$ with $J \in \mathbb{N}$ as

$$\mathrm{mixt}^J\mathcal{Q} = \left\{ \sum_{j=1}^{J} \lambda_j q_j \,|\, \lambda \in \Delta^J,\, q_j \in \mathcal{Q},\, \forall j \in \{1, \dots, J\} \right\}. \tag{39}$$

Distributions from a mixture family can approximate any probability density arbitrarily well [91] provided that the allowed number of components is high enough. This result shows that the family $\mathrm{mixt}\mathcal{Q}$ can approximate any probability density, but it may not be the case of $\mathrm{mixt}^J\mathcal{Q}$ if $J \in \mathbb{N}$ is not high enough. We point to [71] and [53] where the approximation error coming from an insufficient number of mixture components is analysed in terms of the mass-covering and mode-seeking KL divergences.

When the number of components J is fixed, the family $\mathrm{mixt}^J\mathcal{Q}$ can be interpreted as a parametrised family, with parameters being the mixtures weights and the parameters of each components. However, it is often more efficient to leverage the extra structure provided by the mixture form of the distributions.

4.2 Infinite-dimensional approximating families

In the previous section, we have presented parametric families of probability distributions $\mathcal{Q}$ for which each element $p \in \mathcal{Q}$ has a density that can be evaluated, can be sampled from, and is represented by a finite-dimensional parameter. Such families are called explicit in [36]. These features make the families we have seen so far easy to work with. However, we usually have $\mathcal{Q} \subsetneq \mathcal{P}(\mathcal{X})$, meaning that such families may lack expressivity. We discuss in this section the possibility of working within the full space $\mathcal{P}(\mathcal{X})$.

Since distributions from $\mathcal{P}(\mathcal{X})$ will generally be untractable, as it will not be possible to describe them by some finite-dimensional parameters, we will introduce two specific situations that retain some tractability. The first situation corresponds to the case where subsequent distributions remain unavailable but such that it is possible to generate samples from these distributions. Such mechanism is illustrated in Fig. 3. This situation, where probability ditributions are only defined implicitly through their sampling process, is actually quite common, interacting particle systems and Markov chains being the two main examples. Although many works study such processes in continuous time through stochastic differential equations, we try to present discrete-time processes as much as possible since they are closer to implementation.

We will also mention another situation, which consists in initializing dynamics on $\mathcal{P}(\mathcal{X})$ at a probability measure of the form $\frac{1}{N}\sum_{n=1}^{N}\delta_{x_0^n}$. Initializing distributions-valued dynamics at such distributions can have the effect that the subsequent distributions induced by the dynamics will remain sums of Dirac masses. In other words, the dynamics on the space $\mathcal{P}(\mathcal{X})$ is transformed on a dynamics on $\mathcal{X}^N$, and only the Dirac masses locations need to be adapted.

Interacting particles systems and sums of Dirac masses thus can allow for implementable distribution-valued dynamics. However, the end result of such dynamics will be a empirical measure or a sum of Dirac masses, which do not posses a density. While this is perfectly fine for sampling, this can be problematic in situations where the construction of proposals with a density is required. For instance, think of importance sampling, where having a density is required in order to compute importance weights. To cover such situations, we show how to convolutions can be used to construct a density from a probability distribution that may not have one.

4.2.1 Particle systems

We will now present distributions that are constructed by iterating several Markov chain steps. These distributions can be expressed through Markov kernels, they can be sampled from, but they do not necessarily admit a density.

More precisely, consider the sequence of samples $\{x_t\}_{t=0}^{T}$, such that $x_0 \sim p_0$, and for every $t \in [\![1, T-1]\!]$,

$$x_{t+1} \sim P_{t+1}(x_t, \cdot), \tag{40}$$

where each P_t is a Markov kernel (see for instance [25] for more details). Then, the final sample $x_T \sim p_T = P_T \dots P_1 p_0$. Actually, we have at every step $t \in [\![1, T]\!]$ that $x_t \sim P_t \dots P_1 p_0$. Even if p_T can be defined formally, it will generally not admit a density that can be evaluated straightforwardly. However, samples from p_T can be generated by iterating (40) for $t \in [\![1, T-1]\!]$.

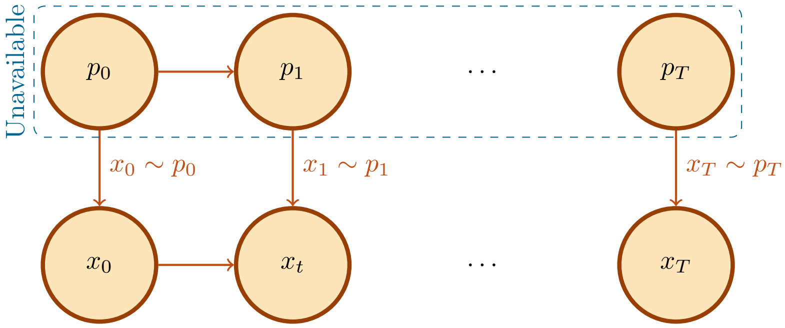


Figure 3: Subsequent samples can be constructed iteratively through stochastic updates. At each iteration, the sample can be considered to be a sample from an underlying probability. Often, it is possible to write a (most of the times, deterministic) dynamics to update the underlying law. When the sequence of probability measures are sequence of an optimisation algorithm to minimise some functional, this framework is often referred under the name *sampling as optimisation*.

As an important example, let us mention that the update (40) can represent Euler discretisations of diffusion processes through

$$x_{t+1} = a_t(x_t) + c_t(x_t)\xi_t, \tag{41}$$

where $\xi_t \sim \mathcal{N}(0, I_d)$. This encompasses for instance unadjusted Langevin Monte Carlo schemes, which are recovered when $a_t = -\nabla \log \pi$ for some probability density π. In this case, the evolution of $\{p_t\}_{t=1}^T$ can be related to Fokker-Planck equations.

One can remark that iterating Markov sampling steps as it is done in (40) is related to the iterated transformations the initial sample undergoes when sampling from normalizing flows as in (37). Actually, the formalism of Markov kernels generalise the normalising flows that we introduced earlier, as it was remarked in [48]. Indeed, a normalizing flow corresponds to a deterministic transform of the current sample, and we can write $x_{t+1} = T_{\theta_t}(x_t)$ as $x_{t+1} \sim \delta_{T_{\theta_t}(x_t)}$, where $\delta_{T_{\theta_t}(\cdot)}(\cdot)$ is a Markov kernel.

In this review, we will not consider that the Markov kernels are parametrized by parameters that can be adapted. There has been some works towards the optimization of these parameters, such as [8, 48] in the context of MCMC algorithms, or [36] in the context of sequential Monte Carlo.

We now present another sampling process, but this time, we will require that several samples are generated at each iteration of the sampling process. This will allow to have interactions between the samples, leading to richer behaviours.

We consider now the sequence $\{\{x_k^n\}_{n=1}^N\}_{t=1}^T$, such that for all $n \in [\![1, N]\!]$, $x_0^n \sim p_0$ and

$$x_{t+1}^n = a_t(x_t^n) + \sum_{m=1}^N b_t(x_t^n, x_t^m) + c_t(x_t^n)\xi_t^n, \; \forall t \in [\![1, T-1]\!], \tag{42}$$

where the $\{\xi_t^n\}_{n=1}^N$ are i.i.d. realisations of some random variable, for instance $\xi_t^n \sim \mathcal{N}(0, I_d)$. Then, the samples $\{x_T^n\}_{n=1}^N$ are i.i.d., and we denote by p_T their probability distribution.

Remark that if $b_k(\cdot,\cdot) \equiv 0$ in (42), meaning that there is no interactions between particles, the N particles will be independent. One can then interpret the dynamics in (42) as N parallel runs from a diffusion-like dynamics as in (41).

Many works have studied the behaviour of the empirical measure $\widehat{p}_t = \frac{1}{N}\sum_{n=1}^N \delta_{x_t^n}$, and shown that the dynamics (42) were related to McKean-Vlasov equations in the same way that diffusions are related to Fokker-Planck equations. The main difference between McKean-Vlasov and Fokker-Planck equations being that in the former, the dynamics itself depends on the current distribution, that is $x_{t+1} = F_t(x_t, p_t)$, while in the latter, $x_{t+1} = F_t(x_t)$. Allowing interacting particles allows to approximate the dependence on p_t by approximating p_t by its empirical measure.

4.2.2 Sum of Dirac masses

So far, we have presented dynamics on $\mathcal{P}(\mathcal{X})$ where the resulting probability distributions $\{p_t\}_{t\geq 0}$ are unavailable or remain implicit but such that it is possible to generate samples $\{\{x_t^n\}_{n=1}^N\}_{t\geq 0}$ whose distributions are the under lying p_t. We now present another situation where dynamics over $\mathcal{P}(\mathcal{X})$ can be implemented exactly. Namely, we discuss here the situation where a distribution-valued dynamics $p_{t+1} = F_t(p_t)$ is initialized at at a sum of Dirac masses of the form $p_0 = \frac{1}{N}\sum_{n=1}^N \delta_{x_0}$.

When a distribution-valued dynamics is initialized at a sum of Dirac masses, then the dynamics can sometimes reduce to a dynamics on the locations of the masses. For instance, consider the update $p_+ = (Id - \nabla V)_{\#}p$. Then, if $p = \frac{1}{N}\sum_{n=1}^N \delta_x^n$, one can compute $p_+ = \frac{1}{N}\sum_{n=1}^N \delta_{x_+^n}$, where for each $n \in [\![1, N]\!]$, we have

$$x_+^n = x^n - \nabla V(x^n).$$

Such a model is used for instance in [53].

4.2.3 Convolutions of empirical measures

We have considered so far probability measures which can be sampled through stochastic dynamics, but whose density is unavailable. In some contexts, this can be a serious drawbacks: importance sampling requires to be able to evaluate the density of the proposal for instance. We now review the use of convolutions to construct a density from samples.

We now define the convolution product. Note that we give a somewhat restricted definition, as we will only consider convolution between a probability density and a probability measure, while it is possible to define the convolution between two measures.

Definition 13. Consider a probability density $k \in \mathcal{P}(\mathcal{X}, m)$, that we will call a kernel, and a probability measure $p \in \mathcal{P}(\mathcal{X})$. Then, the convolution $k * p$ is the probability density in $\mathcal{P}(\mathcal{X}, m)$ defined by

$$(k * p)(x) = \int k(x-y)p(dy), \ \forall x \in \mathcal{X}. \tag{43}$$

We present an important and prototypical example of convolutions. Consider the probability measure $p = \frac{1}{N}\sum_{n=1}^N \delta_{x_n}$, where $N \in \mathbb{N}$ and $x_n \in \mathbb{R}^d$ for every $n \in [\![1, N]\!]$, as well as the centered Gaussian kernel with covariance $\Sigma \in \mathcal{S}_{++}^d$, denoted by k_Σ. Then, we have

$$(k_\Sigma * p)(x) = \frac{1}{N}\sum_{n=1}^N k_\Sigma(x - x_n), \ \forall x \in \mathbb{R}^d.$$

In other words, $k_\Sigma * p$ is the mixture of N Gaussians with covariance Σ and respective means being the samples $\{x_n\}_{n=1}^N$.

Convolutions of the form $k_\Sigma * \left(\frac{1}{N}\sum_{n=1}^N \delta_{x_n}\right)$ also form the basis of kernel density estimators, where one aims to approximate the density of p such that $x_n \sim p$. In order to improve the convergence of $k_\Sigma * \left(\frac{1}{N}\sum_{n=1}^N \delta_{x_n}\right)$ to p, one should have $\Sigma \to 0$ as $N \to +\infty$ (see for instance [91]).

Particle systems with adapted convolution kernels have been used in [73].

5 Different algorithmic principles

Once we have settled for a divergence-minimization formulation ($P_{D,\mathcal{Q}}$) by choosing a divergence D and an approximating family $\mathcal{Q}$, we now need to choose an algorithm to solve the resulting optimization problem. We now review several class of such algorithms below.

Iterative optimisation algorithms generate a sequence of iterates within the search space $\mathcal{Q}$. At every iteration, the next iterate is usually generated by minimising the objective function $D(\pi, \cdot)$, while staying close in some sense to the current iterate. Thus, optimisation algorithms usually assume some kind of geometry on the search space $\mathcal{Q}$. We tried to make this choice as explicit as possible in the upcoming sections, and we focus on three geometries that our the most prevalent among existing algorithms solving ($P_{D,\mathcal{Q}}$).

The three geometries that we focus on are the ones obtained by measuring the proximity between subsequent iterates in the following ways. First, we focus on the case where $\mathcal{Q}$ is a parametric family and where subsequent iterates are compared by computing the Euclidean norm between their parameters. This choice gives rise in particular to the standard gradient descent algorithm. Then, we focus on the case where different proposals in $\mathcal{Q}$ are compared by computing the KL divergence between them. Notable algorithms in this setting are mirror descent and moment-matching algorithms. Finally, we consider the case where distributions in $\mathcal{Q}$ are compared using the Wasserstein distance. This choice allows to consider gradient flows on the space of measures.

5.1 Methods leveraging the geometry induced by the Euclidean distance between parameters

5.1.1 Algorithmic building blocks

In this section, we introduce two optimization operators based on the Euclidean distance between parameters that can be used to perform optimization. These operators are the forward and backward operators. This first one corresponds to the standard gradient descent step and the second one to the proximal operator. When the objective function is a sum of two functions, it is possible to compose an operator associated to the first function with an operator associated to the second function.

Suppose that we have at our disposal a parameter $\theta \in \Theta$ and we wish to minimize the function F which takes parameters $\theta \in \Theta$ as argument.

The Euclidean forward operator associated to F and with step size $\tau > 0$ applied at θ yields θ_+ such that

$$\theta_+ = \theta - \tau \nabla F(\theta). \tag{Eucl-Forward}$$

The dependence on the Euclidean distance may not appear clearly in the above update. Actually, it can be shown that the update in (Eucl-Forward) can be equivalently rewritten as

$$\theta_+ = \operatorname*{arg\,min}_{\theta' \in \Theta} \left\{ F(\theta) + \langle \nabla F(\theta), \theta' - \theta \rangle + \frac{1}{2\tau} \|\theta' - \theta\|^2 \right\}. \tag{44}$$

We can observe that (Eucl-Forward) amounts to the minimization of the sum of a linear approximation of F around θ and a quadratic distance term that penalizes large steps from the current iterate θ.

The Euclidean backward operator, also called the proximal operator, associated to F with step size $\tau > 0$ applied at θ yields θ_+ defined by

$$\theta_+ = \operatorname*{arg\,min}_{\theta' \in \Theta} \left\{ F(\theta') + \frac{1}{2\tau} \|\theta' - \theta\|^2 \right\}. \tag{Eucl-Backward}$$

Compared to the forward operator described in (Eucl-Forward) which used a linear approximation of F around θ, (Eucl-Backward) minimizes the full function F while penalizing large steps from θ. Remark that (Eucl-Backward) does not require F to be differentiable. An intuition of the name of this operator comes from the observation than when F is differentiable, then (Eucl-Backward) is equivalent to

$$\theta_+ = \theta - \tau \nabla F(\theta_+). \tag{45}$$

This expression is very close to the one in (Eucl-Forward) except that the gradient is now taken at θ_+ instead of θ.

The forward and backward Euclidean operators can be combined by splitting the objective function F into $F = F_1 + F_2$ and applying one operator for each term of the sum. For instance, if the backward operator is applied to F_1 after applying the forward operator to F_2 starting from θ, it is possible to show that the updated point θ_+ satisfies

$$\theta_+ = \operatorname*{arg\,min}_{\theta' \in \Theta} \left\{ F_1(\theta') + F_2(\theta) + \langle \nabla F_2(\theta), \theta' - \theta \rangle + \frac{1}{2\tau} \|\theta' - \theta\|^2 \right\}. \tag{46}$$

We call this operator the forward-backward operator.

5.1.2 Implementation and approximation of the gradients

The main challenge in (Eucl-Forward) is usually the computation of the gradient ∇F. Indeed, most statistical divergences of interest involve complicated integrals with respect to the approximating distribution or the target itself. In this context, several strategies have been proposed in the literature.

REINFORCE gradients can be used to estimate the gradients of functions of the form $F(\theta) = \int V(x) q_\theta(x) m(dx)$. Suppose first that one can write $\nabla F(\theta) = \int V(x) \nabla_\theta q_\theta(x) m(dx)$. Then, observe that for a fixed $x \in \mathcal{X}$,

$$\nabla_\theta (\log q_\theta)(x) = \frac{\nabla_\theta q_\theta(x)}{q_\theta(x)}.$$

We conclude from this observation that

$$\nabla F(\theta) = \int V(x) \nabla_\theta (\log q_\theta(x)) q_\theta(x) m(dx).$$

This allows to write $\nabla F(\theta)$ as an expectation with respect to q_θ, and thus to approximate it by

$$\nabla F(\theta) \approx \frac{1}{N} \sum_{n=1}^{N} V(x^n) \nabla_\theta (\log q_\theta(x^n)),$$

where the $\{x^n\}_{n=1}^N$ are N i.i.d. samples from q_θ.

We now present the reparametrized gradient method, which can also be used to approximate gradients of functions of the form $F(\theta) = \int V(x) q_\theta(x) m(dx)$ when q_θ belongs to a reparametrized family. We recall from Section 4.1.3 that q_θ belongs to a reparametrised family if there exists q_0 and T_θ such that if $x \sim q_0$, then $T_\theta(x) \sim q_\theta$. In this case, we can rewrite $F(\theta)$ as

$$F(\theta) = \int V(T_\theta(x)) q_0(dx).$$

In particular, working with a reparametrised family allows to write F as en expectation with respect to a probability distribution that does not depend on θ. Then, ∇F is approximated by

$$\nabla F(\theta) \approx \frac{1}{N} \sum_{n=1}^{N} Jac_\theta T_\theta(x^n) \nabla_x V(T_\theta(x^n)),$$

where for $n \in [\![1, N]\!]$, $x^n \sim q_0$.

We now present the stick-the-landing (STL) estimator, which can be used to compute the gradient of the entropy functional $H(\theta) = \int \log(q_\theta(x))q_\theta(x)m(dx)$.

So far, we have only discussed gradient estimators for functions of θ which are integrals with respect to q_θ. However, we sometimes deal with integrals with respect to the target π. Such integrals arise for instance when minimising the inclusive KL divergence $\theta \longmapsto KL(\pi, q_\theta)$. Usually, we do not have samples from π at hand, so one needs to resort to another approximation procedure. Two main approaches coexist. The first one is based on importance sampling [27, 78, 46] and the second one uses MCMC [84, 59].

Forward-backward updates in [34], to alleviate the lack of smoothness of the entropy term arising when $KL(\cdot, \pi)$ is being minimized.

5.1.3 Literature

- Minimisation of the exclusive KL: [94] for the original paper, see also [35, 59] for convergence results.
- Minimisation of the inclusive KL (using Markov chains): [84, 58]
- Minimisation of the inclusive χ^2 divergence: [4] over exponential families, [81] over normalising flows
- Minimisation of Rényi divergences: [72]

5.2 Methods leveraging the geometry induced by the Kullback-Leibler divergence between proposals

Comparing the proximity between subsequent iterates by measuring the Euclidean distance between their parameters suffers from several drawbacks. First, it does not allow to work with distributions that cannot be described by finite-dimensional parameters. Second, it neglects the facts that all the parameters may not have the same role and that they may be constrained. For instance, the covariance of a Gaussian distribution needs to stay positive-definite. In this section, we will thus compare the proximity between subsequent iterates using the Kullback-Leibler divergence between them.

5.2.1 Building blocks

The Kullback-Leibler divergence is not symmetric. We will thus introduce more operators than usual since the order of the terms in the penalization can also change. We will introduce operators based on parametric families as well as operators working on the full space of probability measures.

Mirror descent operators Backward operators can be readily proposed by changing the squared Euclidean distance in (Eucl-Backward) to a KL divergence. For the forward operators, we will leverage the proximal-like interpretation of (Eucl-Forward) and change the squared Euclidean distance to a KL divergence. We will see that these forward operators can be seen as mirror descent updates.

We start by introducing a KL-based forward operator on the full space of probability measures, which updates q to q_+ such that

$$q_+ = \arg\min_{q'} \left\{ F(q) + \int \frac{\delta F}{\delta q}(q)(x)(q' - q)(dx) + \frac{1}{\tau} KL(q', q) \right\} \tag{$\overleftarrow{\text{KL}}$-Forward}$$

We now propose a similar forward operator that will work for parametric set of approximating distributions. This operator updates θ to θ_+ such that

$$\theta_+ = \arg\min_{\theta' \in \Theta} \left\{ F(\theta) + \langle \nabla F(\theta), \theta' - \theta \rangle + \frac{1}{\tau} KL(q_\theta, q_{\theta'}) \right\}. \tag{$\overrightarrow{\text{KL}}$-Forward}$$

Natural gradient descent Another widely used method to optimize parameters that leverages a KL-based geometry is the natural gradient algorithm. Consider the update in ($\overrightarrow{\text{KL}}$-Forward) which involves the KL divergence between subsequent proposals instead of the Euclidean squared distance between their parameter. In order to tame the non-linearities appearing with the KL divergence, one can expand it around the parameters of one of its arguments and obtain [6]:

$$KL(q_\theta, q_{\theta'}) = KL(q_{\theta'}, q_\theta) = \frac{1}{2}\langle \theta - \theta', I(\theta)(\theta - \theta')\rangle + o(\|\theta - \theta'\|^2). \tag{47}$$

This means that the KL divergence divergence can locally be approximated by a Mahalanobis distance between the parameters with shape matrix given by the Fisher information matrix of q_θ. This motivates us to consider the operator

$$\theta_+ = \arg\min_{\theta' \in \Theta} \left\{ F(\theta) + \langle \nabla F(\theta), \theta' - \theta\rangle + \frac{1}{2\tau}\|\theta' - \theta\|^2_{I(\theta)} \right\}. \tag{48}$$

This update can be then shown to be equivalent to what we will call the natural gradient descent update, that is

$$\theta_+ = \theta - \tau (I(\theta))^{-1}\nabla F(\theta). \tag{Natural-Forward}$$

5.2.2 Implementation

Tempering The density-valued forward operator ($\overleftarrow{\text{KL}}$-Forward) can be shown to admit a closed form (see [29, Equation (2)] or [24, Equation (6)]). Indeed, if q_+ has been obtained as in ($\overleftarrow{\text{KL}}$-Forward), we have that

$$q_+(x) \propto \exp\left(-\tau \frac{\partial F}{\partial q}(q)(x)\right) q(x), \; \forall x \in \mathcal{X}.$$

In the particular case where $F = KL(\cdot, \pi)$, ($\overleftarrow{\text{KL}}$-Forward) can be seen as a form of tempering [24], that is

$$q_+(x) \propto \pi(x)^\tau q(x)^{1-\tau}, \; \forall x \in \mathcal{X}.$$

In this case, q_+ can be seen as a geometric average between the current iterate q and the target distribution π. Recall that tempering has been widely used to create intermediate target distributions that are easier to approximate than π in a variety of contexts [85, 83, 61].

The special case of exponential families Exponential families are particularly well-suited to work within the KL-based geometry, as the operators ($\overrightarrow{\text{KL}}$-Forward) and (Natural-Forward) admit explicit closed forms. These tow operators can also be shown to be dual to each other in some sense. We now make these statements precise using preliminary results presented in Section 4.1.2.

We start by writing ($\overrightarrow{\text{KL}}$-Forward) as a mirror descent update taking place in the space of dual parameters. Indeed, we have from Proposition 4 that $KL(q_\theta, q_{\theta'}) = d_A(\theta', \theta)$. We can thus show by writing the optimality conditions that the forward update ($\overrightarrow{\text{KL}}$-Forward) is equivalent to

$$\nabla A(\theta_+) = \nabla A(\theta) - \tau \nabla F(\theta). \tag{49}$$

Since ∇A is invertible and its inverse is usually straightforwardly available, one just needs to compute $\nabla F(\theta)$ and perform a gradient descent step in the dual parameter space to implement ($\overrightarrow{\text{KL}}$-Forward). These updates have been used in [46, 13] and can be written as relaxed moment-matching updates for some choices of F [46].

We now show that the natural gradient descent update (Natural-Forward) also admits a mirror descent interpretation and admits a straightforward implementation. This mirror descent interpretation has been first proposed in [95] and has been used recently in [105] to analyse the convergence of some stochastic natural gradient variational inference algorithms.

In the case of an exponential family, we have that the Fisher information matrix $I(\theta)$ is equal to the Hessian of the log-partition function, that is, $I(\theta) = \nabla^2 A(\theta)$. Further, since $(\nabla A)^{-1} = \nabla A^*$, we obtain that $\nabla^2 A(\theta)^{-1} = \nabla^2 A^*(\nabla A(\theta))$ and that

$$\theta_+ = \theta - \tau \nabla (F \circ \nabla A^*)(\nabla A(\theta)). \tag{50}$$

If we write $\eta = \nabla A(\theta)$ and $\eta_+ = \nabla A(\theta_+)$, we can write ($\overrightarrow{\mathrm{KL}}$-Forward) as

$$\nabla A^*(\eta_+) = \nabla A^*(\eta) - \tau \nabla (F \circ \nabla A^*)(\eta). \tag{51}$$

One can check that this mirror descent update is equivalent to the proximal-like update

$$\eta_+ = \operatorname*{arg\,min}_{\eta' \in \mathrm{dom}\, A^*} \left\{ (F \circ \nabla A^*)(\eta) + \langle \nabla (F \circ \nabla A^*)(\eta), \eta' - \eta \rangle + \frac{1}{\tau} d_{A^*}(\eta, \eta') \right\}.$$

From these calculations, we gather that in an exponential family, (Natural-Forward) can be implemented without the need to invert a Hessian, which is a costly operation, and that it can be interpreted as a mirror descent update on the function $F \circ \nabla A^*$.

5.2.3 Literature

- natural gradient VI: NGVI [57], for mixture families [74], [105, 63] for some kinds of guarantees, see also [79] about natural gradient in general
- about the infinite-dimensional mirror descent update ($\overleftarrow{\mathrm{KL}}$-Forward) and links with tempering: [29, 61, 24]
- Instances and analysis of ($\overrightarrow{\mathrm{KL}}$-Forward) for α- and Rényi divergences: [46, 13] when the exponential family is used, [30] for more general families (note that the linear approximation of F in ($\overrightarrow{\mathrm{KL}}$-Forward) is replaced with something a bit more intricate but still linked with mirror descent)

5.3 Methods leveraging the geometry induced by the Wasserstein methods between proposals

We now consider algorithms that leverage the geometry induced by the Wasserstein distance. This geometry is well-suited to work over the full space of probability measures and it also exhibits nice properties when working over the set of Gaussian densities. The Wasserstein distance is symmetric, i.e., $W_2(q_1, q_2) = W_2(q_2, q_1)$, hence we do not have different types of updates as it was the case for updates based on the KL divergence. If Euclidean algorithms were standard proximal and gradient descent algorithms and KL-based algorithms were instances of mirror descent algorithms, we will see that Wasserstein-based algorithms can be seen a Riemannian optimization algorithms.

5.3.1 Building blocks

Full space of measures We now present some Wasserstein-based operators that allow to work directly on the full space of probability measures.

The first operator is the Wasserstein forward operator, which can be understood as a gradient descent step on the Wasserstein space (the space of probability measures with second-order moments endowed with the Wasserstein distance W_2). Given a distribution q in this space, the Wasserstein forward operator associated to F with step size τ yields the updated probability distribution q_+ defined as

$$q_+ = (Id - \tau \nabla_W F(q))_\# \, q. \tag{W-Forward}$$

The above update uses the so-called Wasserstein gradient of F at q, which is defined using the first variation of F by

$$\nabla_W F(q)(x) = \nabla \left(\frac{\partial F}{\partial q}(q) \right)(x), \, \forall x \in \mathcal{X}.$$

We now introduce the following backward operator leveraging the Wasserstein distance. Its definition is comparable to the definitions on (Eucl-Backward): this operators updates q so that F is minimized as much as possible while staying close to the starting point q.

$$q_+ = \arg\min_{q'} \left\{ F(q') + \frac{1}{2\tau} W_2(q', q)^2 \right\} \tag{W-Backward}$$

The special case of Gaussian measures The space of Gaussian measures with the Wasserstein distance is called the Bures-Wasserstein space. Actually, when the space of Gaussian measures is endowed with the Wasserstein distance W_2, it forms a submanifolds of the Wasserstein space. This means in particular that the trajectory of geodesics between two Gaussians in the Wasserstein space is made of Gaussian probability distributions. We can thus specialize the Wasserstein-based operators (W-Forward) and (W-Backward) to the Bures-Wasserstein space.

We first introduce the Bures-Wasserstein forward operator associated to F with step size $\tau > 0$. This operator updates q to q_+ using a push-forward that is similar to the one in (W-Forward), except that we now use the Bures-Wasserstein gradient of F.

$$q_+ = (Id - \tau \nabla_{BW} F(q))_\# \, q \tag{BW-Forward}$$

We now introduce the backward operator on the Bures-Wassertein space, which is completely similar to (W-Backward), except that now, the search space in the inner optimization problem is the space of Gaussian measures.

$$q_+ = \arg\min_{q'} \left\{ F(q') + \frac{1}{2\tau} W_2(q', q)^2 \right\}. \tag{BW-Backward}$$

5.3.2 Implementing Wasserstein-based operators

Using Langevin dynamics These updates are related through the JKO scheme and the Fokker-Planck equation [55]. Indeed, the sequence of measures generated by iterating the proximal step converges to the solution of a Fokker-Planck equation when the step size τ goes to zero. Then, the Euler discretisation of the resulting dynamics yields the gradient update.

Implementing Wasserstein optimisation updates over the full space of probability measures When q is an empirical measure, that is, $q = \frac{1}{N} \sum_{n=1}^N \delta_{x^n}$, it is possible that the distribution q_+ obtained after taking a forward step (W-Forward) is of the form $q_+ = \frac{1}{N} \sum_{n=1}^N \delta_{x_+^n}$. Then, one needs to compute the location points meaning that for any $n \in [\![1, N]\!]$, we have

$$x_n^+ = x_n - \tau \nabla_W F(q)(x_n). \tag{52}$$

This situations arises for instance in [53].

Implementing Wasserstein optimisation updates for Gaussian measures The space of Gaussian measures endowed with the Wasserstein distance is often called the Bures-Wasserstein space and benefits from many nice properties.

We now show that the update (BW-Forward) can be written in a closed form. The Bures-Wasserstein gradient of F is the affine mapping defined by

$$\nabla_{BW} F(q)(x) = \left(\int \nabla^2 \left(\frac{\delta F}{\delta q}(q) \right)(x) q(dx) \right)(x - \mu) + \int \nabla \left(\frac{\delta F}{\delta q}(q) \right)(x) q(dx). \tag{53}$$

We now compute the mean μ_+ and covariance Σ_+ of q_+. Suppose that we have $z_0 \sim \mathcal{N}(0, I)$, then $z = \Sigma^{\frac{1}{2}} z_0 + \mu \sim q$ and $z_+ = z - \tau \nabla_{BW} F(q)(z) \sim q_+$. Using (53), we compute that

$$\begin{aligned} z_+ &= \Sigma^{\frac{1}{2}} z_0 + \mu - \tau \left(\left(\int \nabla^2 \left(\frac{\delta F}{\delta q}(q) \right)(x) q(dx) \right) \Sigma^{\frac{1}{2}} z_0 + \int \nabla \left(\frac{\delta F}{\delta q}(q) \right)(x) q(dx) \right) \\ &= \left(Id - \tau \int \nabla^2 \left(\frac{\delta F}{\delta q}(q) \right)(x) q(dx) \right) \Sigma^{\frac{1}{2}} z_0 + \left(\mu - \tau \int \nabla \left(\frac{\delta F}{\delta q}(q) \right)(x) q(dx) \right). \end{aligned}$$

Identifying the above with the fact that $z_+ = \Sigma_+^{\frac{1}{2}} z_0 + \mu_+$, we obtain that

$$\begin{aligned} \mu_+ &= \mu - \tau \int \nabla \left(\frac{\delta F}{\delta q}(q) \right)(x) q(dx), \\ \Sigma_+ &= \left(Id - \tau \int \nabla^2 \left(\frac{\delta F}{\delta q}(q) \right)(x) q(dx) \right) \Sigma \left(Id - \tau \int \nabla^2 \left(\frac{\delta F}{\delta q}(q) \right)(x) q(dx) \right). \end{aligned}$$

We also give the explicit form of the backward operator (BW-Backward) associated to the entropy function $H(q) = \int \log(q(x)) q(dx)$, which was used in [32]. The distribution q_+ obtained by applying (BW-Backward) to H with step size $\tau > 0$ from q with mean μ and covariance Σ is the Gaussian distribution q_+ with mean μ_+ and covariance Σ_+ given by

$$\begin{aligned} \mu_+ &= \mu, \\ \Sigma_+ &= \frac{1}{2} \left(\Sigma + 2\tau Id + (\Sigma(\Sigma + 4\tau Id))^{\frac{1}{2}} \right). \end{aligned}$$

The properties of the Bures-Wasserstein space have been leveraged to minimise the exclusive KL divergence over Gaussian distributions in [65, 32]. The forward operator (BW-Forward) was used in [65], while a forward-backward algorithm was used in [32]. Remark that [65] also proposed an algorithm to minimise $KL(\cdot, \pi)$ over mixtures of Gaussians.

5.4 Methods based on other principles

Other geometries Fisher-Rao geometry, Stein variational gradient descent,...

Laplace method

The population Monte Carlo algorithm The population Monte Carlo algorithm is a popular importance sampling algorithm that also uses convolutions to construct proposals. However, the samples used to perform the convolution are not explicitly following an infinite-dimensional divergence-minimizing procedure.

$$\begin{aligned}
p_k &= \text{Sampling-Resampling procedure} \\
x_k^{(n)} &\sim p_k \\
q_k &= \left(\frac{1}{N} \sum_{n=1}^{N} \delta_{x_k^{(n)}} \right) * K_h .
\end{aligned}$$

The optimization interpretation of the dynamics of $\{p_k\}_{k \in \mathbb{N}}$ are to our knowledge not clearly established.

6 A synthetic view of the algorithms

6.1 A classification based on the choice of approximating family

We now present a wide array of methods that can be shown in some sense to solve an instance of Problem ($P_{D,\mathcal{Q}}$). These methods use different families $\mathcal{Q}$, and we highlight these features in Table 1.

	Gaussian	Exp. families	Reparam. families	Non-parametric
Simple		NGD VI [57]	Black-box VI [94]	Langevin diffusion [55]
		Moment matching [14]	Var. Rényi bound [72]	Tempering [9, 24]
		Optimal AIS [96, 4]	CHIVI [33]	
		Relaxed moment-matching [46]		
Mixture	BW SGD [65]	MCEF [74]		
	GRAMIS [39]	Monotonic α-div. [30]		

Table 1: Showing which approximating family each method uses

All the methods that are shown in Table 1 to use mixture approximating families can also by used with only one single component in the mixture. Usually, the resulting approximating family is less expressive, but the theoretical analysis is generally more thorough in the single component case.

6.2 A classification based on the divergence being minimized

We have seen so far that many problems in statistics boil down to an instance of Problem ($P_{D,\mathcal{Q}}$), and similarly, many existing algorithms are solving instances of Problem ($P_{D,\mathcal{Q}}$). We consider in this section several algorithms from variational inference, adaptive importance sampling, and statistical inference in general. We classify these algorithms in Table 2 based on the divergence that is being minimized.

$KL(\cdot,\pi)$	$KL(\pi,\cdot)$	$D_\alpha(\cdot,\pi)$	$D_\alpha(\pi,\cdot)$
Black-box VI [94] NGD VI [57, 74] BW SGD [65] Langevin diffus. [55] Tempering [9, 24]	Moment matching [14, 27] M-PMC [20]	Var. Rényi bound [72] CHIVI [33]	Monotonic α-div. [30] Relaxed moment matching [46] Optimal AIS [96, 4] Escort moment matching [47]

Table 2: Showing which divergence each method minimizes

Some of the considered algorithms explicitly minimize a divergence, as it is the case for most considered VI algorithms [94, 72, 33, 57, 46, 65, 30, 47] or some AIS algorithms [20, 96, 4]. For some other algorithms, things can be a bit more fuzzy as the link with divergence minimization is not always stated explicitly. This is the case of many methods based on moment-matching, such as [20, 38], which in fact amounts to minimizing the inclusive KL divergence [14]. This is also the case of some procedures or dynamics, such as the Langevin diffusion, which has been recognized to be a minimization algorithm for the exclusive KL divergence in the landmark work of [55]. This is also the case of the tempering procedure, which has been used a lot in computational statistics [85, 83], and has been shown to be a mirror descent algorithm minimizing the exclusive KL divergence [9, 24].

Some algorithms presented in Table 2 minimizing an α-divergence sometimes hold for a range of values of α allowing them to minimize a KL divergence. The monotonic α-divergence minimization algorithm of [30] and the relaxed moment-matching algorithm of [46] can also be used to minimize the inclusive KL divergence, while the variational Rényi bound algorithm of [72] can be used to minimize the exclusive KL divergence.

A Optimization over the space of measures

Definition 14. First variation

Proposition 5. *Show the variational interpretation of Bayesian updates?*

Can be done easily using the machinery of [66].

Corollary 1. *Compute the update* $\gamma^{\textit{excl-KL}}_{\tau,KL(\cdot,\pi)}(q)$.

B The Wasserstein distance and Wasserstein space

Good references are [23] and [89].

References

[1] S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. Stuart. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, 32(3):405–431, 2017.

[2] Y. Akimoto, Y. Nagata, I. Ono, and S. Kobayashi. Theoretical foundation for CMA-ES from information geometry perspective. *Algorithmica*, 64(4):698–716, 2012.

[3] Y. Akimoto and Y. Ollivier. Objective improvement in information-geometric optimization. In *Conference on Foundations of Genetic Algorithms (FOGA)*, 2013.

[4] O. Akyildiz and J. Míguez. Convergence rates for optimised adaptive importance samplers. *Statistics and Computing*, 31(12), 2021.

[5] O. D. Akyildiz, F. R. Crucinio, M. Girolami, T. Johnston, and S. Sabanis. Interacting particle Langevin algorithm for maximum marginal likelihood estimation. *ESAIM: Probability and Statistics*, 2025.

[6] S. Amari. *Differential-Geometrical Methods in Statistics.* Springer New York, 1985.

[7] M. Amrouche, H. Carfantan, and J. Idier. Efficient sampling of Bernoulli-Gaussian-mixtures for sparse signal restoration. *IEEE Transactions on Signal Processing*, 70:5578–5591, 2022.

[8] C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statistics and Computing*, 18:343–373, 2008.

[9] P.-C. Aubin-Frankowski, A. Korba, and F. Léger. Mirror descent with relative smoothness in measure spaces, with application to Sinkhorn and EM. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2022.

[10] Y. Bai, A. B. Dieker, and H. Lam. Curse of dimensionality in rare-event simulation. In *Winter Simulation Conference (WSC)*, pages 375–384, 2023.

[11] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6(58):1705–1749, 2005.

[12] O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory.* John Wiley & Sons, Ltd, 2014.

[13] F. Bertholom, R. Douc, and F. Roueff. Asymptotics of alpha-divergence variational inference algorithmds with exponential families. In *Conference on Neural Information Processins Systems (NeurIPS)*, 2024.

[14] C. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[15] C. Bishop and H. Bishop. *Deep Learning: Foundations and Concepts.* Springer, 2024.

[16] D. Blei, A. Kucukelbir, and J. McAuliffe. Variational inference: A review for the statistician. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[17] L. Brown. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory.* Institute of Mathematical Statistics, 1986.

[18] M. F. Bugallo, V. Elvira, L. Martino, D. Luengo, J. Míguez, and P. M. Djuric. Adaptive importance sampling: The past, the present, and the future. *IEEE Signal Processing Magasine*, 34(4):60–79, 2017.

[19] T. Campbell and B. Beronov. Sparse variational inference: Bayesian coresets from scratch. In *Advances in Neural Information Processing Systems (NeurPIS)*, volume 32, 2019.

[20] O. Cappé, R. Douc, A. Guillin, J. M. Marin, and C. P. Robert. Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18:447–459, 2008.

[21] G. Chantas, N. Galatsanos, A. Likas, and M. Saunders. Variational Bayesian image restoration based on a product of t-distributions image prior. *IEEE Transactions on Image Processing*, 17(10):1795–1805, 2008.

[22] S. Chatterjee and P. Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2), 2018.

[23] S. Chewi, J. Niles-Weed, and P. Rigollet. *Statistical Optimal Transport.* Springer, 2025.

[24] N. Chopin, F. Crucinio, and A. Korba. A connection between tempering and entropic mirror descent. In *International Conference on Machine Learning (ICML)*, 2024.

[25] N. Chopin and O. Papaspiliopoulos. *An Introduction to Sequential Monte Carlo.* Springer, 2020.

[26] A. Cichocki and S.-I. Amari. Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

[27] J. M. Cornuet, J. M. Marin, A. Mira, and C. P. Robert. Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39(4):798–812, 2012.

[28] I. Csiszár and P. Shields. *Information Theory and Statistics: A Tutorial.* Now Foundations and Trends, 2004.

[29] B. Dai, N. He, H. Dai, and L. Song. Provable Bayesian inference via particle mirror descent. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

[30] K. Daudel, R. Douc, and F. Roueff. Monotonic alpha-divergence minimisation for variational inference. *Journal of Machine Learning Research*, 24(62):1–76, 2023.

[31] A. K. Dhaka, A. Catalina, M. Welandawe, M. R. Andersen, J. Huggins, and A. Vehtari. Challenges and opportunities in high dimensional variational inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7787–7798, 2021.

[32] M. Z. Diao, K. Balasubramanian, S. Chewi, and A. Salim. Forward-backward Gaussian variational inference via JKO in the bures-Wasserstein space. In *Internation Conference on Machine Learning (ICML)*, pages 7960–7991, 2023.

[33] A. B. Dieng, D. Tran, R. Ranganath, J. Paisley, and D. Blei. Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

[34] J. Domke. Provable smoothness guarantees for black-box variational inference. In *International Conference on Machine Learning (ICML)*, 2020.

[35] J. Domke, G. Garrigos, and R. Gower. Provable convergence guarantees for black-box variational inference. https://arxiv.org/abs/2306.03638, 2023.

[36] A. Doucet, E. Moulines, and A. Thin. Differentiable samplers for deep latent variable models. *Philosophical Transactions of the Royal Society A*, 381(2247), 2023.

[37] R. L. Dykstra. An iterative procedure for obtaining I-projections onto the intersection of convex sets. *The Annals of Probability*, 13(3):975–984, 1985.

[38] Y. El-Laham, V. Elvira, and M. F. Bugallo. Recursive shrinkage covariance learning in adaptive importance sampling. In *IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 624–628, 2019.

[39] V. Elvira, E. Chouzenoux, O. D. Akyildiz, and L. Martino. Gradient-based adaptive importance samplers. *Journal of the Franklin Institute*, 360(13):9490–9514, 2023.

[40] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Generalized multiple importance sampling. *Statistical Science*, 34(1):129–155, 2019.

[41] C. Fernández and M. F. J. Steel. On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441):359–371, 1998.

[42] T. Geffner and J. Domke. On the difficulty of unbiased alpha divergence minimization. In *International Conference on Machine Learning (ICML)*, 2021.

[43] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su. A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, 2(4):1360–1383, 2008.

[44] P. Grünwald, R. de Heide, and W. M. Koolen. Safe testing. *Journal of the Royal Statistical Society Series B*, 2024.

[45] T. Guilmeau, N. Branchini, E. Chouzenoux, and V. Elvira. Adaptive importance sampling for heavy-tailed distributions via α-divergence minimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

[46] T. Guilmeau, E. Chouzenoux, and V. Elvira. Regularized Rényi divergence minimization through Bregman proximal gradient algorithms. https://arxiv.org/abs/2211.04776, 2022.

[47] T. Guilmeau, E. Chouzenoux, and V. Elvira. On variational inference and maximum likelihood estimation with the λ-exponential family. *Foundations of Data Science*, 2024. https://arxiv.org/2310.05781.

[48] P. Hagemann, J. Hertrich, and G. Steidl. Stochastic normalizing flows for inverse problems: A Markov chains viewpoint. *SIAM/ASA Journal on Uncertainty Quantification*, 2022.

[49] N. Hansen. The CMA evolution strategy: A tutorial. https://arxiv.org/abs/1604.00772, 2023.

[50] N. Hansen, D. V. Arnold, and A. Auger. Evolution strategies. In *Springer Handbook of Computational Intelligence*. Springer, 2015.

[51] T. Hastie, R. Tibshirani, and J. Firedman. *The Elements of Statistical Learning*. Springer, 2009.

[52] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(4):1303–1347, 2013.

[53] T. Huix, A. Korba, A. O. Durmus, and E. Moulines. Theoretical guarantees for variational inference with fixed-variance mixture of Gaussians. In *International Conference on Machine Learning (ICML)*, 2024.

[54] P. Jaini, I. Kobyzev, Y. Yu, and M. Brubaker. Tails of Lipschitz triangular flows. In *International Conference on Machine Learning (ICML)*, 2020.

[55] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker-Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17.

[56] M. Khan and W. Lin. Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 878–887, 2017.

[57] M. E. Khan and D. Nielsen. Fast yet simple natural-gradient descent for variational inference in complex models. In *International Symposium on Information Theory and its Applications (ISITA)*, page 31635, 2018.

[58] K. Kim, J. Oh, J. R. Gardner, A. B. Dieng, and H. Kim. Markov chain score ascent: A unifying framework of variational inference with Markovian gradients. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[59] K. Kim, J. Oh, K. Wu, Y.-A. Ma, and J. R. Gardner. On the convergence of black-box variational inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[60] J. Knoblauch, J. Jewson, and T. Damoulas. An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23(132), 2022.

[61] A. Korba and F. Portier. Adaptive importance sampling meets mirror descent : a bias-variance tradeoff. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151, pages 11503–11527, 2022.

[62] M. A. Kumar and I. Sason. Projection theorems for the Rényi divergence on α-convex sets. *IEEE Transactions on Information Theory*, 62(9):4924–4935, 2016.

[63] N. Kumar, T. Möllenhoff, M. E. Khan, and A. Lucchi. Optimization guarantees for square-root natural-gradient variational inference. *Transactions on Machine Learning Research*, 2025.

[64] J. Kuntz, J. N. Lim, and A. M. Johansen. Particle algorithms for maximum likelihood training of latent variable models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

[65] M. Lambert, S. Chewi, F. Bach, S. Bonnabel, and P. Rigollet. Variational inference via Wasserstein gradient flows. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[66] N. Lanzetti, S. Bolognani, and F. Dörfler. First-order conditions for optimization in the Wasserstein space. *SIAM Journal on Mathematics of Data Science*, 7(1), 2025.

[67] P. Larrañaga. A review on estimation of distribution algorithms. In *Estimation of Distribution Algorithms. A New Tool for Evolutionary Computation.* Springer, 2002.

[68] F. Laus, F. Pierre, and G. Steidl. Nonlocal myriad filters for Cauchy noise removal. *Journal of Mathematical Imaging and Vision*, 60:1324–1354, 2018.

[69] H. Lee, C. Pabbaraju, A. P. Sevekari, and A. Risteki. Universal approximation using well-conditioned normalizing flows. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[70] C. T. Li and F. Farnia. Mode-seeking divergences: Theory and applications to GANs. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 8321–8350, 2023.

[71] J. Q. Li and A. R. Barron. Mixture density estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1999.

[72] Y. Li and R. Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

[73] J. N. Lim and A. Johansen. Particle semi-implicit variational inference. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[74] W. Lin, M. E. Khan, and M. Schmidt. Fast and simple natural-gradient variational inference with mixture of exponential-family approximations. In *International Conference on Machine Learning (ICML)*, 2019.

[75] F. Llorente, L. Martino, D. Delgado, and J. López-Santiago. Marginal likelihood computation for model selection and hypothesis testing: An extensive review. *SIAM Review*, 65:3–58, 2023.

[76] C. Margossian, L. Pillaud-Vivien, and L. Saul. An ordering of divergences for variational inference with factorized gaussian approximations. https://arxiv.org/abs/2403.13748.

[77] C. Margossian and L. Saul. The shrinkage-delinkage trade-off: An analysis of factorized Gaussian approximations for variational inference. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2023.

[78] J.-M. Marin, P. Pudlo, and M. Sedki. Consistency of adaptive importance sampling and recycling schemes. *Bernoulli*, 25(3), 2019.

[79] J. Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 2020.

[80] A. M. Metelli, A. Russo, and M. Restelli. Subgaussian and differentiable importance sampling for off-policy evaluation and learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[81] L. I. Midgey, V. Stimper, G. N. C. Simm, B. Schölkopf, and J. M. Hernandez-Lobato. Flow annealed importance sampling bootstrap. In *International Conference on Learning Representations (ICLR)*, 2023.

[82] T. L. Minh, J. Arbel, T. Moellenhoff, M. E. Khan, and F. Forbes. Natural variational annealing for multimodal optimization. https://arxiv.org/abs/2501.04667, 2025.

[83] P. D. Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society Series B*, 68(3):411–436, 2006.

[84] C. Naesseth, F. Lindsten, and D. M. Blei. Markovian score climbing: Variational inference with KL(p——q). In *Advances in Neural Information Processing Systems (Neurips)*, 2020.

[85] R. M. Neal. Annealing importance sampling. *Statistics and Computing*, 11:125–139, 2001.

[86] F. Nielsen and R. Nock. Entropies and cross-entropies of exponential families. In *IEEE International Conference on Image Processing (ICIP)*, pages 3621–3624, 2010.

[87] Y. Ollivier, L. Arnold, A. Auger, and N. Hansen. Information-geometric optimization algorithms: A unifying picture via invariance principles. *Journal of Machine Learning Research*, 18(18):1–65, 2017.

[88] A. B. Owen. *Monte Carlo theory, methods and examples.* https://artowen.su.domains/mc/, 2013.

[89] V. M. Panaretos and Y. Zemel. *An Invitation to Statistics in Wasserstein Space.* Springer, 2020.

[90] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, (21):1–64, 2021.

[91] E. Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[92] D. Peel and G. J. McLachlan. Robust mixture modelling using the t-distribution. *Statistics and Computing*, 10:339–348, 2000.

[93] Y. Polyanskiy and Y. Wu. Information theory: From coding to learning. https://people.lids.mit.edu/yp/homepage/papers.html, 2023.

[94] R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 33, pages 814–822, 2014.

[95] G. Raskutti and S. Mukherjee. The information geometry of mirror descent. *IEEE transactions on Information Theory*, 61(3):1451–1457, 2015.

[96] E. Ryu and S. Boyd. Adaptive importance sampling via stochastic convex programming. 12 2014.

[97] D. Sanz-Alonso. Importance sampling and necessary sample size: An information theory approach. *SIAM/ASA Journal on Uncertainty Quantification*, 6:867–879, 2018.

[98] M. Teboulle. A simplified view of first order methods for optimization. *Mathematical Programming*, 170(1):67–96, 2018.

[99] Y. Tikochinsky, N. Z. Tishby, and R. D. Levine. Alternative approach to maximum-entropy inference. *Physical Review A*, 30:2638–2644, 1984.

[100] M. E. Tipping and N. D. Lawrence. Variational inference for Student-t models: Robust Bayesian interpolation and generalised component analysis. *Neurocomputing*, 69:123–141, 2005.

[101] T. van Erven and P. Harremoes. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions of Information Theory*, 60(7):3797–3820, 2014.

[102] S. Wang and T. Swartz. Moment matching adaptive importance sampling with skew-Student proposals. *Monte Carlo Methods and Applications*, 28(2):149–162, 2022.

[103] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.

[104] D. Wierstra, T. Schaul, T. Glasmachers, Y. Sun, J. Peters, and J. Schmidhuber. Natural evolution strategies. *Journal of Machine Learning Research*, 15:949–980, 2014.

[105] K. Wu and J. Gardner. Understanding stochastic natural gradient variational inference. In *International Conference on Machine Learning (ICML)*, 2024.

[106] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt. Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):2008–2026, 2019.