q-Paths: Generalizing the Geometric Annealing Path using Power Means

Vaden Masrani[1] **Rob Brekelmans**[2] **Thang Bui**[3] **Frank Nielsen**[4] **Aram Galstyan**[2] **Greg Ver Steeg**[2]
Frank Wood[1,5]

[1]University of British Columbia, Vancouver, Canada
[2] University of Southern California, Information Sciences Institute, Marina del Rey, CA USA
[3]University of Sydney, Sydney, Australia
[4]Sony Computer Science Laboratories Tokyo, Japan
[5]Montreal Institute for Learning Algorithms Montreal, Canada

Abstract

Many common machine learning methods involve the geometric annealing path, a sequence of intermediate densities between two distributions of interest constructed using the geometric average. While alternatives such as the moment-averaging path have demonstrated performance gains in some settings, their practical applicability remains limited by exponential family endpoint assumptions and a lack of closed form energy function. In this work, we introduce q-paths, a family of paths which is derived from a generalized notion of the mean, includes the geometric and arithmetic mixtures as special cases, and admits a simple closed form involving the deformed logarithm function from nonextensive thermodynamics. Following previous analysis of the geometric path, we interpret our q-paths as corresponding to a q-exponential family of distributions, and provide a variational representation of intermediate densities as minimizing a mixture of α-divergences to the endpoints. We show that small deviations away from the geometric path yield empirical gains for Bayesian inference using Sequential Monte Carlo and generative model evaluation using Annealed Importance Sampling.

1 INTRODUCTION

Given a tractable and often normalized base distribution $\pi_0(z)$ and unnormalized target $\tilde{\pi}_1(z)$, many statistical methods require a path $\gamma : [0,1] \to \mathcal{P}$, where $\mathcal{P}$ is a family of unnormalized density functions with $\gamma(0) = \pi_0(z)$ and $\gamma(1) = \tilde{\pi}_1(z)$. For example, marginal likelihood estimation methods such as thermodynamic integration (TI) (Ogata, 1989) or Annealed Importance Sampling (AIS) (Neal, 2001) and Markov Chain Monte Carlo (MCMC) methods such as parallel tempering (Earl and Deem, 2005) and Sequential Monte Carlo (SMC) (Del Moral et al., 2006) typically use the geometric path with mixing parameter β,

$$\tilde{\pi}_\beta(z) = \exp\left\{(1-\beta)\log \pi_0(z) + \beta \log \tilde{\pi}_1(z)\right\}, \qquad (1)$$

In the Bayesian context, $\pi_0(z)$ and $\pi_1(z)$ can represent the prior and posterior distribution, respectively, in which case the geometric path amounts to tempering the likelihood term (Friel and Pettitt, 2008; Nguyen et al., 2015).

Previous work has demonstrated theoretical or empirical improvements upon the geometric path can be achieved, but the applicability of these methods remains limited in practice due to restrictive assumptions on the parametric form of the endpoint distributions. Gelman and Meng (1998) derive an optimal path in distribution space but this is intractable to implement beyond toy examples. The moment-averaging path of Grosse et al. (2013) demonstrates performance gains for partition function estimation in Restricted Boltzmann Machines, but is only applicable for endpoint distributions which come from an exponential family. Bui (2020) proposed a path based on α-divergence minimization using an iterative projection scheme from Minka (2005) which is also reliant on exponential family assumptions.

In this work, we propose q-paths, which can be constructed between arbitrary endpoint distributions and admit a closed form that can be used directly for MCMC sampling

$$\tilde{\pi}_{\beta,q}(z) = \left[(1-\beta)\,\pi_0(z)^{1-q} + \beta\,\tilde{\pi}_1(z)^{1-q}\right]^{\frac{1}{1-q}} \qquad (2)$$

Our q-paths adapt the α-integration of Amari (2007) to the problem of annealing between two unnormalized densities, with our notation q intended to highlight connections with the deformed logarithm and exponential functions from nonextensive thermodynamics (Tsallis, 2009; Naudts, 2011). q-paths may be viewed as taking the generalized mean (Kolmogorov, 1930; de Carvalho, 2016) of the endpoint densities

Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021), PMLR 161:1938–1947.

	Geometric Path	q-Path
Closed Form	$\tilde{\pi}_\beta(z) = \pi_0(z)^{1-\beta}\tilde{\pi}_1(z)^\beta$	$\tilde{\pi}_{\beta,q}(z) = \left[(1-\beta)\,\pi_0(z)^{1-q} + \beta\,\tilde{\pi}_1(z)^{1-q}\right]^{\frac{1}{1-q}}$
Log Linear	$\tilde{\pi}_\beta(z) = \exp\{(1-\beta)\log\pi_0(z) + \beta\log\tilde{\pi}_1(z)\}$	$\tilde{\pi}_{\beta,q}(z) = \exp_q\{(1-\beta)\,\ln_q\pi_0(z) + \beta\,\ln_q\tilde{\pi}_1(z)\}$
Exponential Family	$\tilde{\pi}_\beta(z) = \pi_0(z)\exp\left\{\beta\cdot\log\frac{\tilde{\pi}_1(z)}{\pi_0(z)}\right\}$	$\tilde{\pi}_{\beta,q}(z) = \pi_0(z)\exp_q\left\{\beta\cdot\ln_q\frac{\tilde{\pi}_1(z)}{\pi_0(z)}\right\}$
Variational Representation	$\pi_\beta(z) = \underset{r}{\mathrm{argmin}}(1-\beta)D_{\mathrm{KL}}[r\\|\pi_0] + \beta D_{\mathrm{KL}}[r\\|\pi_1]$	$\tilde{\pi}_{\beta,q}(z) = \underset{\tilde{r}}{\mathrm{argmin}}(1-\beta)D_\alpha[\tilde{\pi}_0\\|\tilde{r}] + \beta D_\alpha[\tilde{\pi}_1\\|\tilde{r}]$

Figure 1: Summary of q-paths (right) in relation to the geometric path (left). q-paths recover the geometric path as $q \to 1$ and $\alpha = 2q-1$ in Amari's α-divergence D_α. The deformed logarithm $\ln_q$ and its inverse $\exp_q$ are defined in Section 2.3.

according to a mixing parameter β and monotonic transformation function $\ln_q(u) = \frac{1}{1-q}(u^{1-q}-1)$. As $q \to 1$, we recover the natural logarithm and geometric mean in Eq. (1), while the arithmetic mean corresponds to $q=0$.

As previous analysis of the geometric path revolves around the exponential family of distributions (Grosse et al., 2013; Brekelmans et al., 2020a,b), we show in Sec. 4 that our proposed paths have an interpretation as a q-exponential family of density functions

$$\tilde{\pi}_{\beta,q}(z) = \pi_0(z)\;\exp_q\left\{\beta\cdot\ln_q\frac{\tilde{\pi}_1(z)}{\pi_0(z)}\right\}. \tag{3}$$

Grosse et al. (2013) show that intermediate distributions along the geometric and moment-averaged paths correspond to the solution of a weighted forward or reverse KL divergence minimization objective, respectively. In Sec. 5, we generalize these variational representations to q-paths, showing that $\tilde{\pi}_{\beta,q}(z)$ minimizes the expected α-divergence to the endpoints for an appropriate mapping between q and α.

Finally, we highlight several implementation considerations in Sec. 7, observing that $q = 1-\delta$ for small δ appears most useful both for qualitative mixing behavior and numerical stability. We provide a simple heuristic for setting an appropriate value of q, and find that q-paths can yield empirical gains for Bayesian inference using SMC and marginal likelihood estimation for generative models using AIS.

2 BACKGROUND

2.1 GEOMETRIC ANNEALING PATH

The geometric mixture path is the most ubiquitous method for specifying a set of intermediate distributions between a tractable base distribution π_0 and unnormalized target $\tilde{\pi}_1$,

$$\pi_\beta(z) = \frac{\pi_0(z)^{1-\beta}\,\tilde{\pi}_1(z)^\beta}{Z(\beta)}, \quad \text{where} \tag{4}$$

$$Z(\beta) = \int \pi_0(z)^{1-\beta}\,\tilde{\pi}_1(z)^\beta dz. \tag{5}$$

The geometric path may also be written as an exponential family of distributions, with natural parameter β and sufficient statistic $T(z) = \log\tilde{\pi}_1(z)/\pi_0(z)$ corresponding to the log importance ratio. We follow Grünwald (2007); Brekelmans et al. (2020a,b) in referring to this as a *likelihood ratio exponential family*, with

$$\pi_\beta(z) = \pi_0(z)\exp\left\{\beta\cdot\log\frac{\tilde{\pi}_1(z)}{\pi_0(z)} - \psi(\beta)\right\} \tag{6}$$

$$\psi(\beta) := \log Z(\beta) = \log\int \pi_0(z)^{1-\beta}\tilde{\pi}_1(z)^\beta dz. \tag{7}$$

It is often more convenient to work with Eq. (6), because one gains access to known exponential family properties that are not apparent from Eq. (4) (Grosse et al., 2013; Brekelmans et al., 2020a,b). In Section 4 we provide an analogous interpretation for q-paths in terms of q-exponential families.

2.2 MOMENT AVERAGING PATH

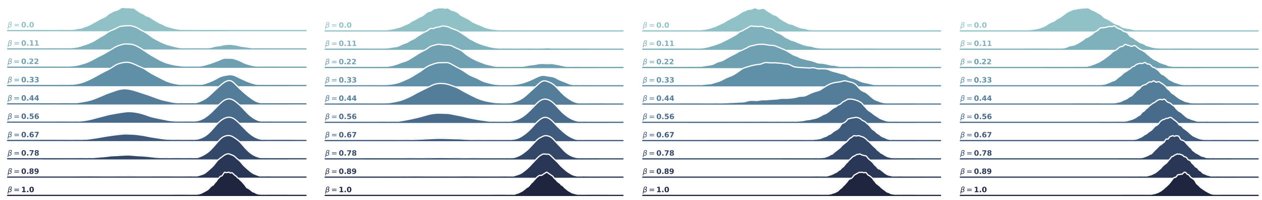
Previous work (Grosse et al., 2013) considers alternative annealing paths in the restricted setting where $\pi_0(z)$ and $\pi_1(z)$ are members of the same exponential family, with parameters θ_0 and θ_1 respectively. Writing the base measure as $g(z)$ and sufficient statistics as $\phi(z)$,

$$\pi_\theta(z) = g(z)\exp\{\theta\cdot\phi(z) - \psi(\theta)\} \tag{8}$$

Grosse et al. (2013) propose the *moment-averaged* path based on the dual or 'moment' parameters of the exponential family, which correspond to the expected sufficient statistics

$$\eta(\theta) = \frac{d\psi(\theta)}{d\theta} = \langle\mathbb{E}_{\pi_\theta}\left[\phi_j(z)\right]\rangle_{j=1}^N, \tag{9}$$

with $\langle\cdot\rangle$ indicating vector notation and $\psi(\theta)$ denoting the log partition function of Eq. (8). In minimal exponential families, the sufficient statistic function $\eta(\theta)$ is a bijective mapping between a natural parameter vector and dual parameter vector (Wainwright and Jordan, 2008).



(a) $q = 0$ (b) $q = 0.5$ (c) $q = 0.9$ (d) $q = 1$

Figure 2: Intermediate q-path densities between $\mathcal{N}(-4, 3)$ and $\mathcal{N}(4, 1)$, with 10 equally spaced β. For low q, the q-path approaches a mixture distribution at $q = 0$, and becomes the geometric mixture parameterized by β at $q = 1$.

The moment-averaged path is defined using a convex combination of the dual parameter vectors (Grosse et al., 2013)

$$\eta(\theta_\beta) = (1-\beta)\,\eta(\theta_0) + \beta\,\eta(\theta_1)\,. \tag{10}$$

To solve for the corresponding natural parameters, we calculate the Legendre transform, or a function inversion η^{-1}.

$$\theta_\beta = \eta^{-1}\big((1-\beta)\,\eta(\theta_0) + \beta\,\eta(\theta_1)\big)\,. \tag{11}$$

This inverse mapping is often not available in closed form and can itself be a difficult estimation problem (Wainwright and Jordan, 2008; Grosse et al., 2013), which limits the applicability of the moment-averaged path in practice.

2.3 Q-DEFORMED LOGARITHM / EXPONENTIAL

While the standard exponential arises in statistical mechanics via the Boltzmann-Gibbs distribution, Tsallis (1988) proposed a generalized exponential which has formed the basis of nonextensive thermodynamics and found wide application in the study of complex systems (Gell-Mann and Tsallis, 2004; Tsallis, 2009).

Consider modifying the integral representation of the natural logarithm $\ln u := \int_1^u \frac{1}{x}dx$ using an arbitrary power function

$$\ln_q u = \int_1^u \frac{1}{x^q}dx. \tag{12}$$

Solving Eq. (12) yields the definition of the q-logarithm

$$\ln_q(u) := \frac{1}{1-q}\left(u^{1-q} - 1\right)\,. \tag{13}$$

We define the q-exponential as the inverse of q-logarithm $\exp_q(u) := \ln_q^{-1}(u)$

$$\exp_q(u) = \big[1 + (1-q)\,u\big]_+^{\frac{1}{1-q}}\,, \tag{14}$$

where $[x]_+ = \max\{0, x\} = \text{RELU}(x)$ ensures that $\exp_q(u)$ is non-negative and fractional powers can be taken for $q < 1$, and thus restricts the domain where $\exp_q(u)$ takes nonzero values to $u > -1/(1-q)$. We omit this notation in subsequent derivations because our q-paths in Eq. (2) take non-negative densities as arguments for the $1/(1-q)$ power.

Note also that both the q-log and q-exponential recover the standard logarithm and exponential function in the limit,

$$\begin{aligned}
&\lim_{q\to 1}\ln_q(u) && \lim_{q\to 1}\exp_q(u)\\
&= \lim_{q\to 1}\frac{\frac{d}{dq}(u^{1-q}-1)}{\frac{d}{dq}(1-q)} && = \lim_{q\to 1}\,[1 + (1-q)\cdot u]^{\frac{1}{1-q}}\\
&= \frac{-\log u \cdot u^{1-q}}{-1}\bigg|_{q=1} && = \lim_{n\to\infty}\left[1 + \frac{u}{n}\right]^n\\
&= \log(u) && := \exp(u).
\end{aligned}$$

In Section 4 we use this property to show q-paths recover the geometric path as $q \to 1$.

3 Q-PATHS FROM POWER MEANS

q-paths are derived using a generalized notion of the mean due to Kolmogorov (1930). For any monotonic function $h(u)$, we define the *generalized mean*

$$\mu_h(\mathbf{u}, \mathbf{w}) = h^{-1}\left(\sum_{i=1}^N w_i \cdot h(u_i)\right), \tag{15}$$

where μ_h outputs a scalar given a normalized measure $\mathbf{w} = (w_1, ..., w_N)$ (with $\sum_{i=1}^N w_i = 1$) over a set of input elements $\mathbf{u} = (u_1, ..., u_N)$ (de Carvalho, 2016).[1]

The generalized mean can be thought of as first applying a nonlinear transformation function to each input, applying the desired weights in the transformed space, and finally mapping back to the distribution space.

The geometric and arithmetic means are *homogeneous*, that is, they have the linear scale-free property $\mu_h(c \cdot \mathbf{u}, \mathbf{w}) = c \cdot \mu_h(\mathbf{u}, \mathbf{w})$. Hardy et al. (1953) shows the unique class of functions $h(u)$ that yield means with the

[1]The generalized mean is also referred to as the *abstract*, *quasi-arithmetic*, or *Kolmogorov-Nagumo* mean in the literature.

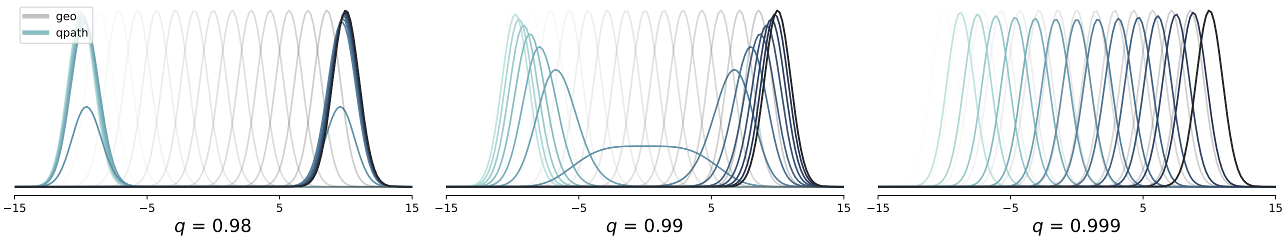


Figure 3: q-paths between $\mathcal{N}(-10, 1)$ and $\mathcal{N}(10, 1)$, which are notably more separated than those in Fig. 2. For difficult annealing problems such as those in our experiments, small deviations from the geometric path (grey) can achieve mass-covering behaviour (center), which is lost if the q-path too much resembles the arithmetic (left) or geometric mean (right).

homogeneity property are of the form

$$h_q(u) = \begin{cases} a \cdot u^{1-q} + b & q \neq 1 \\ \log u & q = 1 \end{cases}. \tag{16}$$

for any a and b. Setting $a = b = 1/(1-q)$, we can recognize $h_q(u)$ as the deformed logarithm $\ln_q(u)$ from Eq. (13).

Generalized means which use the class of functions $h_q(u)$ we refer to as *power means*, and show in App. A that for any choice of a and b,

$$\mu_{h_q}(\mathbf{u}, \mathbf{w}) = \left[\sum_{i=1}^{N} w_i \cdot u_i^{1-q}\right]^{\frac{1}{1-q}}. \tag{17}$$

Notable examples include the arithmetic mean at $q = 0$, geometric mean as $q \to 1$, and the $\min$ or $\max$ operation as $q \to \pm\infty$. For $q = \frac{1+\alpha}{2}$, $a = \frac{1}{1-q}$, and $b = 0$, the function $h_q(u)$ matches the α-representation in information geometry (Amari, 2016), and the resulting power mean over normalized probability distributions as input $\mathbf{u}$ is known as the α-integration (Amari, 2007).

For annealing between unnormalized density functions, we propose the q-path of intermediate $\tilde{\pi}_{\beta,q}(z)$ based on the power mean. Observing that the geometric mixture path in Eq. (1) takes the form of a generalized mean for $h(u) = \ln(u)$, we choose the deformed logarithm

$$h_q(u) := \ln_q(u) \qquad h_q^{-1}(u) = \exp_q(u), \tag{18}$$

as the transformation function for the power mean. This choice will facilitate our parallel discussion of geometric and q-paths in terms of generalized logarithms and exponentials in Section 4.

Using $\mathbf{u} = (\pi_0, \tilde{\pi}_1)$ as the input elements and $\mathbf{w} = (1-\beta, \beta)$ as the mixing weights in Eq. (17), we obtain a simple, closed form expression for the q-path intermediate densities

$$\tilde{\pi}_{\beta,q}(z) = \left[(1-\beta)\,\pi_0(z)^{1-q} + \beta\,\tilde{\pi}_1(z)^{1-q}\right]^{\frac{1}{1-q}} \tag{19}$$

Crucially, Eq. (19) can be directly used as an energy function in MCMC sampling methods such as Hamiltonian Monte Carlo (HMC) (Neal, 2011), and our q-paths do not require additional assumptions on the endpoint distributions.

Finally, to compare against the geometic path, we write the q-path in terms of the generalized mean in Eq. (15)

$$\tilde{\pi}_{\beta,q} = \exp_q\left\{(1-\beta)\,\ln_q \pi_0(z) + \beta\,\ln_q \tilde{\pi}_1(z)\right\}, \tag{20}$$

from which we can see that $\tilde{\pi}_{\beta,q}$ recovers the geometric path in Eq. (1) as $q \to 1$, $\ln_q(u) \to \log(u)$, and $\exp_q(u) \to \exp(u)$. Taking the deformed logarithm of both sides also yields an interpretation of the geometric or q-paths as $\ln$ or $\ln_q$-mixtures of density functions, respectively.

4 Q-LIKELIHOOD RATIO EXPONENTIAL FAMILIES

Similarly to Eq. (6), we relate $\tilde{\pi}_{\beta,q}$ to a q-exponential family with a single sufficient statistic and natural parameter β

$$\tilde{\pi}_{\beta,q}(z) = \left[(1-\beta)\pi_0(z)^{1-q} + \beta\tilde{\pi}_1(z)^{1-q}\right]^{\frac{1}{1-q}} \tag{21}$$

$$= \left[\pi_0(z)^{1-q} + \beta\left(\tilde{\pi}_1(z)^{1-q} - \pi_0(z)^{1-q}\right)\right]^{\frac{1}{1-q}} \tag{22}$$

$$= \pi_0(z)\left[1 + \beta\left(\left(\frac{\tilde{\pi}_1(z)}{\pi_0(z)}\right)^{1-q} - 1\right)\right]^{\frac{1}{1-q}} \tag{23}$$

$$= \pi_0(z)\left[1 + (1-q)\,\beta \ln_q\left(\frac{\tilde{\pi}_1(z)}{\pi_0(z)}\right)\right]^{\frac{1}{1-q}} \tag{24}$$

$$= \pi_0(z) \exp_q\left\{\beta \cdot \ln_q\left(\frac{\tilde{\pi}_1(z)}{\pi_0(z)}\right)\right\}. \tag{25}$$

To mirror the likelihood ratio exponential family interpretation of the geometric path in Eq. (6), we multiply by a factor

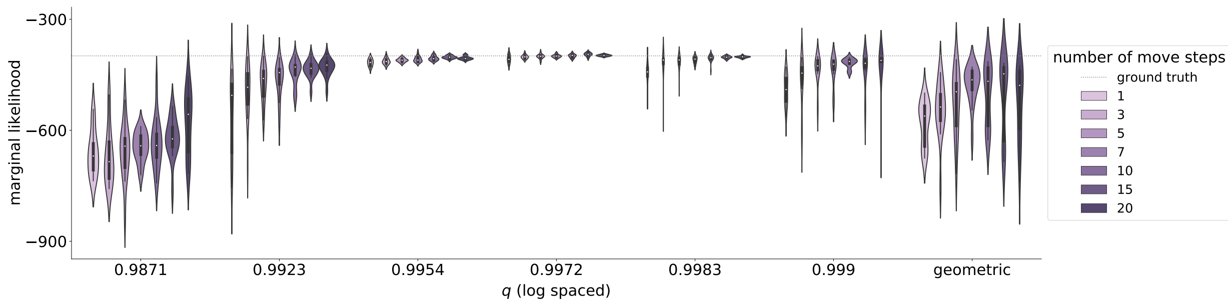


Figure 4: SMC tempering using q-Paths on a binary regression model over 10 runs (cf. Appendix G)). $q = 0.9972$ outperforms the geometric path both in terms of marginal likelihood estimation and reduced variability across runs.

$Z_q(\beta)$ to write the normalized q-path distribution as

$$\pi_{\beta,q}(z) = \frac{1}{Z_q(\beta)}\,\pi_0(z)\exp_q\{\beta\cdot T(z)\} \tag{26}$$

$$Z_q(\beta) := \int \tilde{\pi}_{\beta,q}(z)\,dz\,,\quad T(z) := \ln_q \frac{\tilde{\pi}_1(z)}{\pi_0(z)} \tag{27}$$

which recovers Eq. (6) as $q \to 1$.

Note that we normalize using $Z_q(\beta)$ instead of subtracting a $\psi_q(\beta)$ term inside the $\exp_q$ as in the standard definition of a parameteric q-exponential family (Naudts, 2009, 2011; Amari and Ohara, 2011)

$$\pi_{\theta,q}(z) = g(z)\,\exp_q\left\{\theta\cdot\phi_q(z) - \psi_q(\theta)\right\}. \tag{28}$$

where we use $\phi_q(z)$ to indicate a general sufficient statistic vector which may differ from $T(z) = \ln_q \tilde{\pi}_1(z)/\pi_0(z)$ above.

While $\log Z(\beta) = \psi(\beta)$ for $q = 1$, translating between these normalization constants for $q \neq 1$ requires a non-linear transformation of the parameters. This delicate issue of normalization has been noted in (Matsuzoe et al., 2019; Suyari et al., 2020; Naudts, 2011), and we give detailed discussion in App. B. In App. D, we use the $\psi_q(\theta)$ normalization constant to derive an analogue of the moment-averaging path between parametric q-exponential family endpoints.

q-Paths for Parametric Endpoints The geometric path has a particularly simple form when annealing between exponential family endpoint distributions

$$\theta_\beta = (1-\beta)\,\theta_0 + \beta\,\theta_1\,. \tag{29}$$

In Appendix D.2, we verify Eq. (29) and show that the same result holds for q-paths between endpoint distributions within the same q-exponential family. Intuitively, for the (generalized) exponential family distribution in Eq. (28), we can write the unnormalized density ratio $\ln_q \tilde{\pi}_\theta(z)/g(z) = \theta\cdot\phi(z)$ as a linear function of the parameters θ. Thus, the q-path generalized mean over density functions with $h_q(\tilde{\pi}_{\theta_i}) = \ln_q \tilde{\pi}_{\theta_i}(z)$ will translate to an arithmetic mean in the parameter space with $h_1(\theta_i) = \theta_i$.

5 VARIATIONAL REPRESENTATIONS

Grosse et al. (2013) observe that intermediate distributions along the geometric path can be viewed as the solution to a weighted KL divergence minimization

$$\pi_\beta = \underset{r}{\operatorname{argmin}}(1-\beta)D_{\mathrm{KL}}[r\|\pi_0] + \beta D_{\mathrm{KL}}[r\|\pi_1] \tag{30}$$

where the optimization is over arbitrary distributions $r(z)$.

When the endpoints come from an exponential family of distributions and the optimization is limited to only this parametric family $\mathcal{P}_e$, Grosse et al. (2013) find that the moment-averaged path is the solution to a KL divergence minimization with the order of the arguments reversed

$$\pi_\eta = \underset{r\in\mathcal{P}_e}{\operatorname{argmin}}(1-\beta)D_{\mathrm{KL}}[\pi_0\|r] + \beta D_{\mathrm{KL}}[\pi_1\|r]. \tag{31}$$

In App. C, we follow similar derivations as Amari (2007) to show that the q-path density $\tilde{\pi}_{\beta,q}$ minimizes the α-divergence to the endpoints

$$\tilde{\pi}_{\beta,q} = \underset{\tilde{r}}{\operatorname{argmin}}(1-\beta)D_\alpha[\tilde{\pi}_0||\tilde{r}] + \beta D_\alpha[\tilde{\pi}_1||\tilde{r}] \tag{32}$$

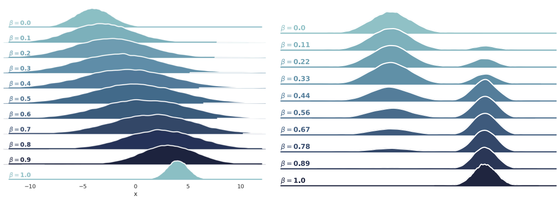
where the optimization is over arbitrary measures $\tilde{r}(z)$. Amari's α-divergence over unnormalized measures, for $\alpha = 2q-1$ (Amari (2016) Ch. 4), is defined

$$D_\alpha[\tilde{r}:\tilde{p}] = \frac{4}{(1-\alpha^2)}\Big(\frac{1-\alpha}{2}\int \tilde{r}(z)dz + \frac{1+\alpha}{2}\int \tilde{p}(z)dz - \int \tilde{r}(z)^{\frac{1-\alpha}{2}}\,\tilde{p}(z)^{\frac{1+\alpha}{2}}dz\Big) \tag{33}$$

The α-divergence variational representation in Eq. (32) generalizes Eq. (30), since the KL divergence $D_{\mathrm{KL}}[\tilde{r}||\tilde{p}]$ is recovered (with the order of arguments reversed)[2] as $q \to 1$.

However, while the α-divergence tends to $D_{\mathrm{KL}}[\tilde{p}||\tilde{r}]$ as $q \to 0$, Eq. (32) *does not* generalize Eq. (31) since the optimization in Eq. (31) is restricted to the parametric family $\mathcal{P}_e$.

[2]The KL divergence extended to unnormalized measures is defined $D_{KL}[\tilde{q}:\tilde{p}] = \int \tilde{q}(z)\log\frac{\tilde{q}(z)}{\tilde{p}(z)}dz - \int \tilde{q}(z)dz + \int \tilde{p}(z)dz$.



(a) Moment-Avg (b) $q = 0$

Figure 5: Moment-averaging path and $q = 0$ mixture path between $\mathcal{N}(-4, 3)$ and $\mathcal{N}(4, 1)$. See Section 5 and Appendix C.1 for discussion.

For the case of arbitrary endpoints, the *mixture* distribution rather than the moment-averaging distribution minimizes the reverse KL divergence in Eq. (31), producing different paths as seen in Fig. 5. We discuss this distinction in greater detail in Appendix C.1.

6 RELATED WORK

In Section 4 and Appendix D, we discuss connections between q-paths and the q-exponential family. Examples of parametric q-exponential families include the Student-t distribution, which has the same first- and second-moment sufficient statistics as the Gaussian and a degrees of freedom parameter ν that specifies a value of $q > 1$. This induces heavier tails than the standard Gaussian and leads to conjugate Bayesian interpretations in hypothesis testing with finite samples (Murphy, 2007; Gelman et al., 2013). The generalized Pareto distribution is another member of the q-exponential family, and has been used for modeling heavy-tail behavior (Pickands III et al., 1975; Bercher and Vignat, 2008; Tsallis, 2009), smoothing outliers for importance sampling estimators (Vehtari et al., 2015), or evaluating variational inference (Yao et al., 2018). q-logarithms and exponentials have also appeared in methods for classification (Ding et al., 2011; Amid et al., 2019), robust hypothesis testing (Qin and Priebe, 2017), mixture modeling (Qin and Priebe, 2013), variational inference (Ding et al., 2011; Kobayashi, 2020), and expectation propagation (Futami et al., 2017; Minka, 2004).

In Section 5, we showed that each q-path density $\tilde{\pi}_{\beta,q}(z)$ specifies the minimizing argument for a variational objective in Eq. (30) or Eq. (32). The value of the objective in Eq. (30) is a mixture of KL divergences, and can be interpreted as a generalized Jensen-Shannon divergence (Nielsen, 2019) or Bregman information (Banerjee et al., 2005). Deasy et al. (2021) explores this mixture of divergences as a regularizer in variational inference, while Brekelmans et al. (2020b) provides additional analysis for case of $q = 1$.

Table 1: SMC sampling with linear/adaptive scheduling in a binary regression model for $\{1, 3, 5\}$ move steps. LIN indicates a linearly spaced schedule ($K = 10$) and ADA uses an adaptive schedule (cf. Section 7.1). Median ERR $= |\log \hat{p}(D) - \log p(D)|$ across 10 seeds is reported against ground truth. Q-PATH (GRID) shows best of 20 log-spaced $\delta \in [10^{-5}, 10^{-1}]$, and Q-PATH (ESS) uses the ESS heuristic to initialize q as described in G.1. Error for most runs (8/12) is Q-PATH (GRID) < Q-PATH (ESS) < GEO.

PIMA	GEO	Q-PATH (ESS HEURISTIC)	Q-PATH (GRID)
LIN-1	79.02 (39.1)	80.64 (42.33)	**10.77 (2.30)**
LIN-3	59.11 (41.71)	59.64 (47.41)	**5.79 (1.46)**
LIN-5	45.63 (19.86)	41.96 (25.23)	**6.63 (2.62)**
ADA-1	2.51 (1.35)	2.31 (2.99)	**1.62 (1.79)**
ADA-3	1.49 (0.43)	1.12 (1.05)	**0.84 (0.84)**
ADA-5	**0.48 (0.60)**	0.76 (0.29)	0.52 (0.59)
SONAR			
LIN-1	228.7 (80.9)	217.92 (72.51)	**93.33 (15.79)**
LIN-3	175.21 (38.66)	172.66 (61.55)	**55.94 (5.69)**
LIN-5	218.94 (92.08)	222.07 (78.76)	**36.67 (10.32)**
ADA-1	20.17 (15.99)	18.15 (15.43)	**15.32 (8.19)**
ADA-3	3.83 (3.44)	3.78 (2.77)	**3.11 (3.26)**
ADA-5	2.79 (2.41)	2.68 (1.95)	**2.23 (0.72)**

7 EXPERIMENTS

Code for all experiments is available at `https://github.com/vmasrani/qpaths_uai_2021`.

7.1 SEQUENTIAL MONTE CARLO IN BAYESIAN INFERENCE

In this section, we use SMC to sample posterior parameters $\pi_1(\theta) = p(\theta|\mathcal{D}) \propto p(\theta)\prod_{n=1}^{N} p(x_n|\theta)$ and estimate the log marginal likelihood $\log p(\mathcal{D}) = \log \int p(\theta)p(\mathcal{D}|\theta)d\theta$ in a Bayesian logistic regression models on the "tall" Pima Indians diabetes dataset ($N = 768, D = 8$) and "wide" Sonar dataset ($N = 208, D = 61$) (see Appendix G). Ground truth $\log p(D)$ is computed using 50k samples and 20 move steps, and for all runs we use 10k samples and plot median error across ten seeds. Grid search shows best of 20 runs, where we sweep over 20 log-spaced $\delta \in [10^{-5}, 10^{-1}]$.

We explore the use of q-paths in both the non-adaptive case, with a fixed linear β schedule with $K = 10$ intermediate distributions, and the adaptive case, where the next value of β_{t+1} is chosen to yield an effective sample size (ESS) of $N/2$ (Chopin and Papaspiliopoulos, 2020).

For the non-adaptive case, we find in Fig. 4 that $q \in [0.9954, 0.9983]$ can achieve more accurate marginal likelihood estimates than the geometric path with fewer movement steps and drastically reduced variance. In Table 1 we see that q-paths achieve gains over the geometric path in

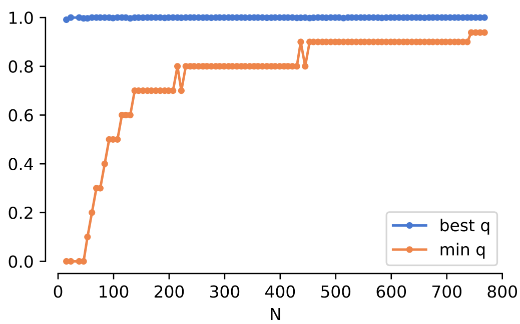


Figure 6: Evaluating the choice of q for SMC. Since the scale of the likelihood $\tilde{\pi}_1$ depends on the number of data examples, we expect the numerical stability of q-paths to vary by N. While the minimum q yielding a stable estimator (orange) increases with N, the best performing q-path (blue) is still $q = 1 - \delta$ for small $\delta > 0$.

both the linear and adaptive setting across both datasets.

Numerical Stability and Implementation To implement q-paths in practice, we begin by considering the log of the expression in Eq. (25), which is guaranteed to be non-negative because $\tilde{\pi}_{\beta,q}(z)$ is an unnormalized density.

$$\log \tilde{\pi}_{\beta,q}(z) = \tag{34}$$
$$\log \pi_0(z) + \frac{1}{1-q} \log \left[1 + (1-q) \cdot \beta \cdot \ln_q \left(\frac{\tilde{\pi}_1(z)}{\pi_0(z)} \right) \right],$$

We focus attention on $\ln_q \tilde{\pi}_1(z)/\pi_0(z)$ term, which is potentially unstable for $q \neq 1$ since it takes importance weights $w = \tilde{\pi}_1(z)/\pi_0(z)$ as input. Since we are usually given log weights in practice, we consider the identity mapping $w = \exp(\log w)$ and reparameterize $q = 1 - \frac{1}{\rho}$ to obtain

$$\ln_q (\exp \log w) = \frac{1}{1-q} \left[(\exp \log w)^{1-q} - 1 \right] \tag{35}$$
$$= \rho \left[(\exp \log w)^{\frac{1}{\rho}} - 1 \right] \tag{36}$$
$$= \rho \left[\exp\{ \frac{1}{\rho} \log w \} - 1 \right]. \tag{37}$$

This suggests q should be chosen such that the exponential doesn't overflow or underflow, which can be accomplished by setting ρ on the order of

$$\rho = \max_i |\log w_i|. \tag{38}$$

where i indexes a set of particles $\{z_i\}$. This choice is reminiscent of the log-sum-exp trick and ensures $|\frac{1}{\rho} \log w| \leq 1$.

In Fig. 6, we explore the impact of changing the scale of $\log w$ on the numerical stability of q-paths. For the case of inferring global model parameters over N i.i.d. data points $p(\mathcal{D}) = \prod_{n=1}^{N} p(x_n)$, we can see that the scale of the unnormalized densities $\tilde{\pi}_1(\theta, \mathcal{D}) = p(\theta) \prod_{n=1}^{N} p(x_n|\theta)$ differs based on the number of datapoints, where increasing N decreases the magnitude of $\log w = \log \tilde{\pi}_1(\theta, \mathcal{D})$ with $\tilde{\pi}_0(\theta) = p(\theta)$.

We randomly subsample N data points for conditioning our model, and observe the effect on both the best-performing q and the numerical stability of SMC with q-paths. The minimum value of q for which we can obtain stable estimators rises as the number of datapoints N increases and the scale of $\tilde{\pi}_1(\theta, \mathcal{D})$ becomes smaller.

Sensitivity to q While setting ρ on the order of $\max_i |\log w_i|$ ensures numeric stability, Fig. 6 indicates that numerical stability may not be sufficient for achieving strong performance in SMC. In fact, q-paths with values just less than 1 consistently perform best across all values of N.

To understand this observation, recall the example in Fig. 3 where the initial and target distribution are well-separated and even the $q = 0.98$ path begins to resemble a mixture distribution. This is clearly undesirable for path sampling techniques, where the goal is to bridge between base and target densities with distributions that are easier to sample.

Heuristic for Choosing q Motivated by the observations above and the desire to avoid grid search, we provide a rough heuristic to find a q which is well-suited to a given estimation problem.

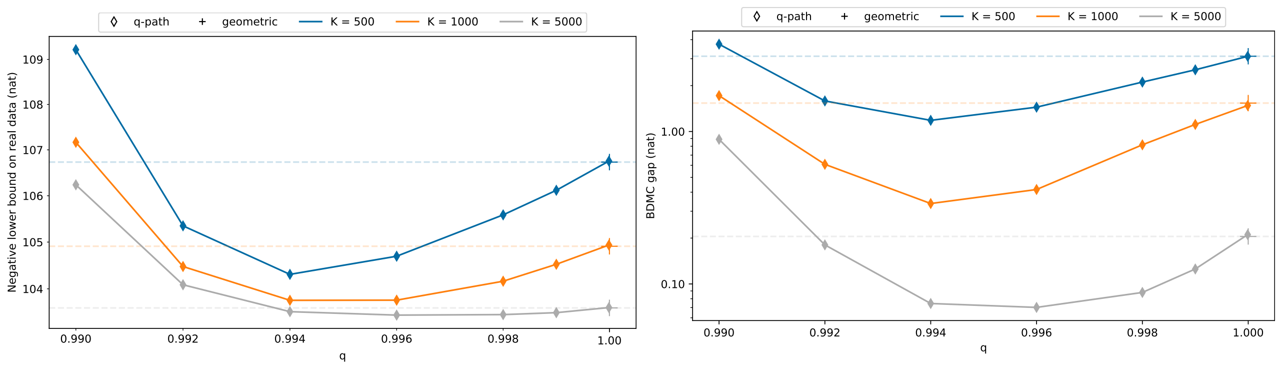
Taking inspiration from the ESS criterion used to select β_{t+1} in our SMC experiments above (Chopin and Papaspiliopoulos, 2020), we select q to obtain a target value of ESS for the first intermediate β_1

$$\mathcal{L}(\beta_1, q) = ||\text{ESS}(\beta_1, q) - \text{ESS}_{\text{target}}||_2^2 \tag{39}$$
$$\text{ESS}(\beta, q) = \frac{\left(\sum_i w_i(\beta, q)\right)^2}{\sum_i w_i(\beta, q)^2} \text{ with } w_i(\beta, q) = \frac{\tilde{\pi}_{\beta,q}(z_i)}{\pi_0(z_i)}.$$

As in the case of the adaptive β scheduling heuristic for SMC, we set the target $\text{ESS}_{\text{target}} = N/2$ to ensure adequate sampling diversity (Jasra et al., 2011; Schäfer and Chopin, 2013; Buchholz et al., 2021; Chopin and Papaspiliopoulos, 2020). For fixed scheduling, the value of β_1 may be known and thus we can easily select q to obtain the target value $\text{ESS}(\beta_1, q) \approx \text{ESS}_{\text{target}}$. However, in adaptive scheduling, β_1 is not known and the objective $\mathcal{L}(\beta_1, q)$ is non-convex in β_1, q. In Appendix G.2, we provide a coordinate descent algorithm to find local optima using random initializations around an initial $q = 1 - \frac{1}{\rho}$ for ρ as in Eq. (38), with results in Table 1.

Note that this heuristic sets q based on a set of initial $z_i \sim \pi_0(z)$, and thus does not consider information about the MCMC sampling used to transform and improve samples.

Nevertheless, in Table 1 we observe that q-paths initial-



(a): Estimating $\log p(x)$ on real data using AIS.

(c): BDMC Gap on simulated data.

Figure 7: Evaluating Generative Models using AIS with q-paths on Omniglot dataset. Best viewed in color.

ized by this heuristic can outperform the geometric path on benchmark SMC binary regression tasks. Comparison with grid search results indicate that further performance gains might be achieved with an improved heuristic.

7.2 EVALUATING GENERATIVE MODELS USING AIS

AIS with geometric paths is often considered the gold-standard for evaluating decoder-based generative models (Wu et al., 2017). In this section, we evaluate whether q-paths can improve marginal likelihood estimation for a variational autoencoder (VAE) trained using the thermodynamic variational objective (TVO) (Masrani et al., 2019) on the Omniglot dataset.

First, we use AIS to evaluate the trained generative model on the true test set, with a Gaussian prior $\pi_0(z) = p(z)$ as the base distribution and true posterior $\pi_1(z) = p(z|x) \propto p(x, z)$ as the target. Intermediate distributions then become $\tilde{\pi}_\beta(z) = p(z)p(x|z)^\beta$. We report stochastic lower bound estimates (Grosse et al., 2015) of $\mathbb{E}_{p_{\text{data}}(x)} \log p(x)$ in Fig. 6(c), where we have plotted the negative likelihood bound so that lower is better. Even for a large number of intermediate distributions, we find that $q \in [0.992, 0.998]$ can outperform the geometric path.

When exact posterior samples are available, we can use a reverse AIS chain from the target density to the base to obtain a stochastic *upper bound* on the log marginal likelihood (Grosse et al., 2015). While such samples are not available on the real data, we can use simulated data drawn from the model using ancestral sampling $x, z \sim p(z)p(x|z)$ as the dataset, and interpret z as a posterior sample. We use the Bidirectional Monte Carlo (BDMC) gap, or difference between the stochastic lower and upper bounds obtained from forward and reverse chains on simulated data, to evaluate the quality of the AIS procedure.

In Fig. 7, we report the average BDMC gap on 2500 simulated data examples, and observe that q-paths with $q = 0.994$ or $q = 0.996$ consistently outperform the geometric path as we vary the number of intermediate distributions K.

8 CONCLUSION

In this work, we proposed q-paths as a generalization of the geometric mixture path which can be constructed between arbitrary endpoint distributions and admits a closed form energy function. We provided a q-likelihood ratio exponential family interpretation of our paths, and derived a variational representation of q-path intermediate densities as minimizing the expected α-divergence to the endpoints. Finally, we observed empirical gains in SMC and AIS sampling using q-paths with $q = 1 - \delta$ for small δ.

Future work might consider more involved heuristics for choosing q, such as running truncated, parallel sampling chains, to capture the interplay between choices of β, q, and sampling method. Applying q-paths in settings such as sampling with parallel tempering (PT) or variational inference using the TVO, remain interesting questions for future work.

Author Contributions

Vaden Masrani and Rob Brekelmans had an equal contribution to this paper.

Acknowledgements

RB and GV acknowledge support from the Defense Advanced Research Projects Agency (DARPA) under awards FA8750-17-C-0106 and W911NF-16-1-0575. VM and FW acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada CIFAR AI Chairs Program, and the Intel Parallel Computing Centers program. This material is based upon work supported by the United States Air Force Research Laboratory (AFRL) under the Defense Advanced Research Projects Agency (DARPA) Data Driven Discovery Models (D3M) program (Contract No. FA8750-19-2-0222) and Learning with Less Labels (LwLL) program (Contract No.FA8750-19-C-0515). Additional support was provided by UBC's Composites Research Network (CRN), Data Science Institute (DSI) and Support for Teams to Advance Interdisciplinary Research (STAIR) Grants. This research was enabled in part by technical support and computational resources provided by WestGrid (https://www.westgrid.ca/) and Compute Canada (www.computecanada.ca).

References

Shun-ichi Amari. Integration of stochastic models by minimizing α-divergence. *Neural computation*, 19(10):2780–2796, 2007.

Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

Shun-ichi Amari and Atsumi Ohara. Geometry of q-exponential family of probability distributions. *Entropy*, 13(6):1170–1185, 2011.

Ehsan Amid, Manfred K Warmuth, and Sriram Srinivasan. Two-temperature logistic regression based on the tsallis divergence. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2388–2396. PMLR, 2019.

Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with Bregman Divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

J-F Bercher and Christophe Vignat. A new look at q-exponential distributions via excess statistics. *Physica A: Statistical Mechanics and its Applications*, 387(22): 5422–5432, 2008.

Rob Brekelmans, Vaden Masrani, Frank Wood, Greg Ver Steeg, and Aram Galstyan. All in the exponential family: Bregman duality in thermodynamic variational inference. In *International Conference on Machine Learning*, 2020a.

Rob Brekelmans, Frank Nielsen, Alireza Makhzani, Aram Galstyan, and Greg Ver Steeg. Likelihood ratio exponential families. *NeurIPS Workshop on Deep Learning through Information Geometry*, 2020b. URL https://openreview.net/forum?id=RoTADibt26_.

Alexander Buchholz, Nicolas Chopin, and Pierre E. Jacob. Adaptive Tuning of Hamiltonian Monte Carlo Within Sequential Monte Carlo. *Bayesian Analysis*, -1(-1):1–27, January 2021. ISSN 1936-0975, 1931-6690. doi: 10.1214/20-BA1222.

Thang Bui. Connecting the thermodynamic variational objective and annealed importance sampling. 2020. URL https://thangbui.github.io/docs/reports/tvo_annealed_is.pdf.

Nicolas Chopin and Omiros Papaspiliopoulos. *An introduction to sequential Monte Carlo*. Springer, 2020.

Miguel de Carvalho. Mean, what do you mean? *The American Statistician*, 70(3):270–274, 2016.

Jacob Deasy, Nikola Simidjievski, and Pietro Liò. Constraining variational inference with geometric jensen-shannon divergence, 2021.

Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

Nan Ding, Yuan Qi, and Svn Vishwanathan. t-divergence based approximate inference. *Advances in Neural Information Processing Systems*, 24:1494–1502, 2011.

David J Earl and Michael W Deem. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005.

Nial Friel and Anthony N Pettitt. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3): 589–607, 2008.

Futoshi Futami, Issei Sato, and Masashi Sugiyama. Expectation propagation for t-exponential family using q-algebra. In *Advances in Neural Information Processing Systems*, pages 2245–2254, 2017.

Murray Gell-Mann and Constantino Tsallis. *Nonextensive entropy: interdisciplinary applications*. Oxford University Press on Demand, 2004.

Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185, 1998.

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

Roger B Grosse, Chris J Maddison, and Ruslan R Salakhutdinov. Annealing between distributions by averaging moments. In *Advances in Neural Information Processing Systems*, pages 2769–2777, 2013.

Roger B Grosse, Zoubin Ghahramani, and Ryan P Adams. Sandwiching the marginal likelihood using bidirectional Monte Carlo. *arXiv preprint arXiv:1511.02543*, 2015.

Peter D Grünwald. *The minimum description length principle*. MIT press, 2007.

G.H. Hardy, J.E. Littlewood, and G. Pólya. Inequalities. *The Mathematical Gazette*, 37(321):236–236, 1953. doi: 10.1017/S0025557200027455.

Ajay Jasra, David A. Stephens, Arnaud Doucet, and Theodoros Tsagaris. Inference for Lévy-Driven Stochastic Volatility Models via Adaptive Sequential Monte Carlo. *Scandinavian Journal of Statistics*, 38(1):1–22, 2011. ISSN 1467-9469. doi: 10.1111/j.1467-9469.2010.00723.x.

Taisuke Kobayashi. q-vae for disentangled representation learning and latent dynamical systems. *arXiv preprint arXiv:2003.01852*, 2020.

Andrey Kolmogorov. On the notion of mean. *Mathematics and Mechanics*, 1930.

Vaden Masrani, Tuan Anh Le, and Frank Wood. The thermodynamic variational objective. *Advances in Neural Information Processing Systems*, 2019.

Hiroshi Matsuzoe, Antonio M Scarfone, and Tatsuaki Wada. Normalization problems for deformed exponential families. In *International Conference on Geometric Science of Information*, pages 279–287. Springer, 2019.

Thomas Minka. Power ep. *Dep. Statistics, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep*, 2004.

Tom Minka. Divergence measures and message passing. Technical report, Microsoft Research, 2005.

Kevin P Murphy. Conjugate bayesian analysis of the gaussian distribution. *def*, 1($2\sigma 2$):16, 2007.

Jan Naudts. The q-exponential family in statistical physics. *Open Physics*, 7(3):405–413, 2009.

Jan Naudts. *Generalised thermostatistics*. Springer Science & Business Media, 2011.

Radford M Neal. Annealed importance sampling. *Statistics and computing*, 11(2):125–139, 2001.

Radford M Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, page 113, 2011.

Thi Le Thu Nguyen, Francois Septier, Gareth W. Peters, and Yves Delignon. Efficient Sequential Monte-Carlo Samplers for Bayesian Inference. *arXiv:1504.05753 [stat]*, April 2015.

Frank Nielsen. On the jensen–shannon symmetrization of distances relying on abstract means. *Entropy*, 21 (5):485, May 2019. ISSN 1099-4300. doi: 10.3390/e21050485. URL `http://dx.doi.org/10.3390/e21050485`.

Yosihiko Ogata. A monte carlo method for high dimensional integration. *Numerische Mathematik*, 55(2):137–157, 1989.

James Pickands III et al. Statistical inference using extreme order statistics. *the Annals of Statistics*, 3(1):119–131, 1975.

Yichen Qin and Carey E Priebe. Maximum l q-likelihood estimation via the expectation-maximization algorithm: a robust estimation of mixture models. *Journal of the American Statistical Association*, 108(503):914–928, 2013.

Yichen Qin and Carey E Priebe. Robust hypothesis testing via l q-likelihood. *Statistica Sinica*, pages 1793–1813, 2017.

Christian Schäfer and Nicolas Chopin. Sequential Monte Carlo on large binary sampling spaces. *Statistics and Computing*, 23(2):163–184, March 2013. ISSN 1573-1375. doi: 10.1007/s11222-011-9299-z.

Hiroki Suyari, Hiroshi Matsuzoe, and Antonio M Scarfone. Advantages of q-logarithm representation over q-exponential representation from the sense of scale and shift on nonlinear systems. *The European Physical Journal Special Topics*, 229(5):773–785, 2020.

Constantino Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of statistical physics*, 52(1-2): 479–487, 1988.

Constantino Tsallis. *Introduction to nonextensive statistical mechanics: approaching a complex world*. Springer Science & Business Media, 2009.

Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*, 2015.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Yuhuai Wu, Yuri Burda, Ruslan Salakhutdinov, and Roger B. Grosse. On the quantitative analysis of decoder-based generative models. In *5th International Conference on Learning Representations, 2017*, 2017.

Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Yes, but did it work?: Evaluating variational inference. In *International Conference on Machine Learning*, pages 5581–5590, 2018.