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

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,[2]USC Information Sciences Institute
[3]University of Sydney, [4]Sony CSL,[5]MILA, [*]Equal Contribution
{vadmas,fwood}@cs.ubc.ca, {brekelma,galstyan,gregv}@isi.edu,
thang.bui@sydney.edu.au, frank.nielsen@acm.org

A ABSTRACT MEAN IS INVARIANT TO AFFINE TRANSFORMATIONS

In this section, we show that $h_q(u)$ is invariant to affine transformations. That is, for any choice of a and b,

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

yields the same expression for the abstract mean μ_{h_α}. First, we note the expression for the inverse $h_q^{-1}(u)$ at $q \neq 1$

$$h_q^{-1}(u) = \left(\frac{u-b}{a}\right)^{\frac{1}{1-q}}. \tag{2}$$

Recalling that $\sum_i w_i = 1$, the abstract mean then becomes

$$\mu_{h_q}(\{w_i\}, \{u_i\}) = h_q^{-1}\left(\sum_i w_i h_q(u_i)\right) \tag{3}$$

$$= h_q^{-1}\left(a\left(\sum_i w_i u_i^{1-q}\right) + b\right) \tag{4}$$

$$= \left(\sum_i w_i u_i^{1-q}\right)^{\frac{1}{1-q}} \tag{5}$$

which is independent of both a and b.

B NORMALIZATION IN Q-EXPONENTIAL FAMILIES

The q-exponential family can also be written using the q-free energy $\psi_q(\theta)$ for normalization Amari and Ohara (2011); Naudts (2011),

$$\pi_{\theta,q}(z) = \pi_0(z)\ \exp_q\left\{\theta \cdot \phi(z) - \psi_q(\theta)\right\}. \tag{6}$$

However, since $\exp_q\{x+y\} = \exp_q\{y\} \cdot \exp_q\{\frac{x}{1+(1-q)y}\}$ (see Suyari et al. (2020) or App. F below) instead of $\exp\{x+y\} = \exp\{x\} \cdot \exp\{y\}$ for the standard exponential, we can not easily move between these ways of writing the q-family Matsuzoe et al. (2019).

Supplement for the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021).

Mirroring the derivations of Naudts (2011) pg. 108, we can rewrite (6) using the above identity for $\exp_q\{x+y\}$, as

$$\pi_\theta^{(q)}(z) = \pi_0(z)\ \exp_q\{\theta\cdot\phi(z) - \psi_q(\theta)\} \tag{7}$$

$$= \pi_0(z)\ \exp_q\{-\psi_q(\theta)\}\exp_q\left\{\frac{\theta\cdot\phi(z)}{1+(1-q)(-\psi_q(\theta))}\right\} \tag{8}$$

Our goal is to express $\pi_\theta^{(q)}(z)$ using a normalization constant $Z_\beta^{(q)}$ instead of the q-free energy $\psi_q(\theta)$. While the exponential family allows us to freely move between $\psi(\theta)$ and $\log Z_\theta$, we must adjust the natural parameters (from θ to β) in the q-exponential case. Defining

$$\beta = \frac{\theta}{1+(1-q)(-\psi_q(\theta))} \tag{9}$$

$$Z_\beta^{(q)} = \frac{1}{\exp_q\{-\psi_q(\theta)\}} \tag{10}$$

we can obtain a new parameterization of the q-exponential family, using parameters β and multiplicative normalization constant $Z_\beta^{(q)}$,

$$\pi_{\beta,q}(z) = \frac{1}{Z_\beta^{(q)}}\pi_0(z)\ \exp_q\{\beta\cdot\phi(z)\} \tag{11}$$

$$= \pi_0(z)\ \exp_q\left\{\theta\cdot\phi(z) - \psi_q(\theta)\right\} = \pi_\theta^{(q)}(z)\,. \tag{12}$$

See Matsuzoe et al. (2019), Suyari et al. (2020), and Naudts (2011) for more detailed discussion of normalization in deformed exponential families.

C MINIMIZING α-DIVERGENCES

Amari (2007) shows that the α power mean $\pi_\beta^{(\alpha)}$ minimizes the expected divergence to a single distribution, for *normalized* measures and $\alpha = 2q-1$. We repeat similar derivations for the case of unnormalized endpoints $\{\tilde\pi_i\}$ and $\tilde r(z)$ and show

$$\tilde\pi_{\beta,q} = \operatorname*{argmin}_{\tilde r(z)}(1-\beta)D_\alpha[\tilde\pi_0(z)||\tilde r(z)] + \beta D_\alpha[\tilde\pi_1(z)||\tilde r(z)], \tag{13}$$

for $\alpha = 2q-1$.

Proof. Defining $w_0 = (1-\beta)$ and $w_1 = \beta$, we consider minimizing the functional

$$r^*(z) = \operatorname*{argmin}_{\tilde r(z)} J[r(z)] = \operatorname*{argmin}_{\tilde r(z)}\left(\sum_{i=0}^{N=1} w_i D_\alpha(\tilde\pi_i(z)||\tilde r(z))\right) \tag{14}$$

Eq. (14) can be minimized using the Euler-Lagrange equations or using the identity

$$\frac{\delta f(x)}{\delta f(x')} = \delta(x-x') \tag{15}$$

from Meng (2004). We compute the functional derivative of $J[r(z)]$ using (15), set to zero and solve for r:

$$\frac{\delta J[r(z')]}{\delta r(z)} = \frac{\delta}{\delta r(z)}\left(\sum_{i=0}^{N=1} w_i\left(\frac{1}{q}\int\tilde p(z')dz + \frac{1}{1-q}\int\tilde r(z')dz - \frac{1}{q(q-1)}\int\tilde\pi_i(z')^{1-q}r(z')^q dz'\right)\right) \tag{16}$$

$$= \left(\sum_{i=0}^{N=1} w_i\left(\frac{1}{1-q}\int\frac{\delta\tilde r(z')}{\delta r(z)}dz - \frac{1}{q(q-1)}\int\tilde\pi_i(z')^{1-q}\cdot q\cdot r(z')^{q-1}\frac{\delta\tilde r(z')}{\delta r(z)}dz'\right)\right) \tag{17}$$

$$= \left(\sum_{i=0}^{N=1} w_i\left(\frac{1}{1-q}\int\delta(z-z')dz - \frac{1}{q-1}\int\tilde\pi_i(z')^{1-q}\cdot r(z')^{q-1}\delta(z-z')dz'\right)\right) \tag{18}$$

$$0 = \frac{1}{1-q} \sum_{i=0}^{N=1} w_i \left(1 - \tilde{\pi}_i(z)^{1-q} \cdot r(z)^{q-1}\right) \tag{19}$$

$$\sum_{i=0}^{N=1} w_i = \sum_{i=0}^{N=1} w_i \tilde{\pi}_i(z)^{1-q} \cdot r(z)^{q-1} \tag{20}$$

$$1 = \sum_{i=0}^{N=1} w_i \tilde{\pi}_i(z)^{1-q} \cdot r(z)^{q-1} \tag{21}$$

$$r(z)^{1-q} = \sum_{i=0}^{N=1} w_i \tilde{\pi}_i(z)^{1-q} \tag{22}$$

$$r(z) = \left[(1-\beta)\tilde{\pi}_0(z)^{1-q} + \beta\tilde{\pi}_1(z)^{1-q}\right]^{1/1-q} = \tilde{\pi}_{\beta,q}(z) \tag{23}$$

□

This result is similar to a general result about Bregman divergences in Banerjee et al. (2005) Prop. 1. although D_α is not a Bregman divergence over normalized distributions.

C.1 ARITHMETIC MEAN ($q = 0$)

For normalized distributions, we note that the moment-averaging path from Grosse et al. (2013) is not a special case of the α-integration Amari (2007). While both minimize a convex combination of reverse KL divergences, Grosse et al. (2013) minimize within the constrained space of exponential families, while Amari (2007) optimizes over all normalized distributions.

More formally, consider minimizing the functional

$$J[r] = (1-\beta)D_{\mathrm{KL}}[\pi_0(z)||r(z)] + \beta D_{\mathrm{KL}}[\pi_1(z)||r(z)] \tag{24}$$

$$= (1-\beta)\int \pi_0(z)\log\frac{\pi_0(z)}{r(z)}dz + \beta\int \pi_1(z)\log\frac{\pi_1(z)}{r(z)}dz \tag{25}$$

$$= \text{const} - \int \left[(1-\beta)\pi_0(z) + \beta\pi_1(z)\right]\cdot \log r(z) dz \tag{26}$$

We will show how Grosse et al. (2013) and Amari (2007) minimize (26).

Solution within Exponential Family Grosse et al. (2013) constrains $r(z) = \frac{1}{Z(\theta)}h(z)\exp(\theta^T g(z))$ to be a (minimal) exponential family model and minimizes (26) w.r.t r's natural parameters θ (cf. Grosse et al. (2013) Appendix 2.2):

$$\theta_i^* = \operatorname*{argmin}_\theta J(\theta) \tag{27}$$

$$= \operatorname*{argmin}_\theta \left(-\int \left[(1-\beta)\pi_0(z) + \beta\pi_1(z)\right]\left[\log h(z) + \theta^T g(z) - \log Z(\theta)\right] dz\right) \tag{28}$$

$$= \operatorname*{argmin}_\theta \left(\log Z(\theta) - \int \left[(1-\beta)\pi_0(z) + \beta\pi_1(z)\right]\theta^T g(z) dz + \text{const}\right) \tag{29}$$

where the last line follows because $\pi_0(z)$ and $\pi_1(z)$ are assumed to be correctly normalized. Then to arrive at the moment averaging path, we compute the partials $\frac{\partial J(\theta)}{\partial \theta_i}$ and set to zero:

$$\frac{\partial J(\theta)}{\partial \theta_i} = \mathbb{E}_r[g_i(z)] - (1-\beta)\,\mathbb{E}_{\pi_0}[g_i(z)] - \beta\,\mathbb{E}_{\pi_1}[g_i(z)] = 0 \tag{30}$$

$$\mathbb{E}_r[g_i(z)] = (1-\beta)\,\mathbb{E}_{\pi_0}[g_i(z)] - \beta\,\mathbb{E}_{\pi_1}[g_i(z)] \tag{31}$$

where we have used the exponential family identity $\frac{\partial \log Z(\theta)}{\partial \theta_i} = \mathbb{E}_{r_\theta}[g_i(z)]$ in the first line.

General Solution Instead of optimizing in the space of minimal exponential families, Amari (2007) instead adds a Lagrange multiplier to (26) and optimizes r directly (cf. Amari (2007) eq. 5.1 - 5.12)

$$r^* = \operatorname*{argmin}_r J'[r] \tag{32}$$

$$= \operatorname*{argmin}_r J[r] + \lambda \left(1 - \int r(z)dz\right) \tag{33}$$

We compute the functional derivative of $J'[r]$ using (15) and solve for r:

$$\frac{\delta J'[r]}{\delta r(z)} = -\int \big[(1-\beta)\pi_0(z') + \beta\pi_1(z')\big] \frac{1}{r(z')} \frac{\delta r(z')}{\delta r(z)} dz' - \lambda \int \frac{\delta r(z')}{\delta r(z)} dz' \tag{34}$$

$$= -\int \big[(1-\beta)\pi_0(z') + \beta\pi_1(z')\big] \frac{1}{r(z')} \delta(z-z')dz' - \lambda \int \delta(z-z')dz' \tag{35}$$

$$= -\big[(1-\beta)\pi_0(z) + \beta\pi_1(z)\big] \frac{1}{r(z)} - \lambda = 0 \tag{36}$$

Therefore

$$r(z) \propto \big[(1-\beta)\pi_0(z) + \beta\pi_1(z)\big], \tag{37}$$

which corresponds to our q-path at $q = 0$, or $\alpha = -1$ in Amari (2007). Thus, while both Amari (2007) and Grosse et al. (2013) start with the same objective, they arrive at different optimum because they optimize over different spaces.

D q-EXPONENTIAL FAMILIES AND ESCORT MOMENT-AVERAGING PATH

In this section, we provide examples of parametric q-exponential family distributions and additional analysis for the special case of annealing between endpoints within the same parametric family. After reviewing the q-Gaussian and Student-t distributions as standard examples of the q-exponential family, we present the *escort*-moments path, which is analogous to Grosse et al. (2013) and relies on the dual parameters of the q-family. We experimentally evaluate these paths in toy examples in Fig. 3, but note that the applicability of the escort-moments path is limited in practice.

D.1 EXAMPLES OF PARAMETRIC q-EXPONENTIAL FAMILY DISTRIBUTIONS

q-Gaussian and Student-t The q-Gaussian distribution appears throughout nonextensive thermodynamics (Naudts, 2009, 2011; Tsallis, 2009), and corresponds to simply taking the $\exp_q$ of the familiar first and second moment sufficient statistics. In what follows, we ignore the case of $q < 1$ since the q-Gaussian has restricted support based on the value of q. For $q > 1$, the q-Gaussian matches the Student-t distribution, whose degrees of freedom parameter ν specifies the order of the q-exponential and introduces heavy tailed behavior.

The Student-t distribution appears in hypothesis testing with finite samples, under the assumption that the sample mean follows a Gaussian distribution. In particular, the degrees of freedom parameter $\nu = n - 1$ can be shown to correspond to an order of the q-exponential family with $\nu = (3-q)/(q-1)$ (in 1-d), so that the choice of q is linked to the amount of data observed.

We can first write the multivariate Student-t density, specified by a mean vector μ, covariance $\mathbf{\Sigma}$, and degrees of freedom parameter ν, in d dimensions, as

$$t_\nu(x|\mu, \mathbf{\Sigma}) = \frac{1}{Z(\nu, \mathbf{\Sigma})} \big[1 + \frac{1}{\nu}(x-\mu)^T \mathbf{\Sigma}^{-1}(x-\mu)\big]^{-\left(\frac{\nu+d}{2}\right)} \tag{38}$$

where $Z(\nu, \mathbf{\Sigma}) = \Gamma(\frac{\nu+d}{2})/\Gamma(\frac{\nu}{2}) \cdot |\mathbf{\Sigma}|^{-1/2} \nu^{-\frac{d}{2}} \pi^{-\frac{d}{2}}$. Note that $\nu > 0$, so that we only have positive values raised to the $-(\nu+d)/2$ power, and the density is defined on the real line.

The power function in (38) is already reminiscent of the q-exponential, while we have first and second moment sufficient statistics as in the Gaussian case. We can solve for the exponent, or order parameter q, that corresponds to $-(\nu+d)/2$ using $-\left(\frac{\nu+d}{2}\right) = \frac{1}{1-q}$. This results in the relations

$$\nu = \frac{d - dq + 2}{q-1} \qquad \text{or} \qquad q = \frac{\nu + d + 2}{\nu + d} \tag{39}$$

We can also rewrite the $\nu^{-1}\,(x-\mu)^T\boldsymbol{\Sigma}^{-1}(x-\mu)$ using natural parameters corresponding to $\{x, x^2\}$ sufficient statistics as in the Gaussian case (see, e.g. Matsuzoe and Wada (2015) Example 4).

Note that the Student-t distribution has heavier tails than a standard Gaussian, and reduces to a multivariate Gaussian as $q \to 1$ and $\exp_q(u) \to \exp(u)$. This corresponds to observing $n \to \infty$ samples, so that the sample mean and variance approach the ground truth (Murphy, 2007).

Pareto Distribution The q-exponential family can also be used for modeling the *tail* behavior of a distribution (Bercher and Vignat, 2008; Vehtari et al., 2015), or, in other words, the probability of $p(x)$ restricted to $X > x_{\text{min}}$ and normalized.

For example, the generalized Pareto distribution is defined via the tail function

$$P(X > x) = \begin{cases} \left(1 + \xi \frac{x - x_{\text{min}}}{\sigma}\right)^{-\frac{1}{\xi}} & \xi \neq 0 \\ \exp\{-\frac{x - x_{\text{min}}}{\sigma}\} & \xi = 0 \end{cases} \tag{40}$$

When $\xi \geq 0$, the domain is restricted to $x \geq x_{\text{min}}$, whereas when $\xi < 0$, the support is between $x_{\text{min}} \leq x \leq x_{\text{min}} - \frac{\sigma}{\xi}$. Writing the CDF as $1 - P(X > x)$ and differentiating leads to

$$p(x) = \frac{1}{\sigma}\left[1 + \xi \cdot \frac{x - x_{\text{min}}}{\sigma}\right]^{-\frac{1}{\xi} - 1} \tag{41}$$

Solving $-\frac{1}{\xi} - 1 = \frac{1}{1-q}$ in the exponent, we obtain $q = \frac{2\xi+1}{\xi+1}$ or $\xi = \frac{q-1}{q-2}$.

D.2 q-PATHS BETWEEN ENDPOINTS IN A PARAMETRIC FAMILY

If the two endpoints $\pi_0, \tilde{\pi}_1$ are within a q-exponential family, we can show that each intermediate distribution along the q-path of the same order is also within this q-family. However, we cannot make such statements for general endpoint distributions, members of different q-exponential families, or q-paths which do not match the index of the endpoint q-parametric families.

Exponential Family Case We assume potentially vector valued parameters $\theta = \{\theta\}_{i=1}^N$ with multiple sufficient statistics $\phi(z) = \{\phi_i(z)\}_{i=1}^N$, with $\theta \cdot \phi(z) = \sum_{i=1}^N \theta_i \phi_i(z)$. For a common base measure $g(z)$, let $\pi_0(z) = g(z)\,\exp\{\theta_0 \cdot \phi(z)\}$ and $\tilde{\pi}_1(z) = g(z)\,\exp\{\theta_1 \cdot \phi(z)\}$. Taking the geometric mixture,

$$\begin{aligned} \tilde{\pi}_\beta(z) &= \exp\left\{(1-\beta)\,\log \pi_0(z) + \beta\,\log \tilde{\pi}_1(z)\right\} && (42)\\ &= \exp\left\{\log g(z) + (1-\beta)\,\theta_0 \cdot \phi(z) + \beta\,\theta_1 \phi(z)\right\} && (43)\\ &= g(z)\exp\left\{\left((1-\beta)\,\theta_0 + \beta\,\theta_1\right) \cdot \phi(z)\right\} && (44) \end{aligned}$$

which, after normalization, will be a member of the exponential family with natural parameter $(1-\beta)\,\theta_0 + \beta\,\theta_1$.

q-Exponential Family Case For a common base measure $g(z)$, let $\pi_0(z) = g(z)\,\exp_q\{\theta_0 \cdot \phi(z)\}$ and $\tilde{\pi}_1(z) = g(z)\,\exp_q\{\theta_1 \cdot \phi(z)\}$. The q-path intermediate density becomes

$$\begin{aligned} \tilde{\pi}_\beta^{(q)}(z) &= \left[(1-\beta)\,\pi_0(z)^{1-q} + \beta\,\tilde{\pi}_1(z)^{1-q}\right]^{\frac{1}{1-q}} && (45)\\ &= \left[(1-\beta)\,g(z)^{1-q}\,\exp_q\{\theta_0 \cdot \phi(z)\}^{1-q} + \beta\,g(z)^{1-q}\,\exp_q\{\theta_1 \cdot \phi(z)\}^{1-q}\right]^{\frac{1}{1-q}} && (46)\\ &= \left[g(z)^{1-q}\left((1-\beta)\,\left[1 + (1-q)(\theta_0 \cdot \phi(z))\right]^{\frac{1}{1-q}1-q} + \beta\,\left[1 + (1-q)(\theta_1 \cdot \phi(z))\right]^{\frac{1}{1-q}1-q}\right)\right]^{\frac{1}{1-q}} \\ &= g(z)\left[1 + (1-q)\left(\left((1-\beta)\,\theta_0 + \beta\,\theta_1\right) \cdot \phi(z)\right)\right]^{\frac{1}{1-q}} && (47)\\ &= g(z)\exp_q\left\{\left((1-\beta)\,\theta_0 + \beta\,\theta_1\right) \cdot \phi(z)\right\} && (48) \end{aligned}$$

which has the form of an unnormalized q-exponential family density with parameter $(1-\beta)\,\theta_0 + \beta\,\theta_1$.

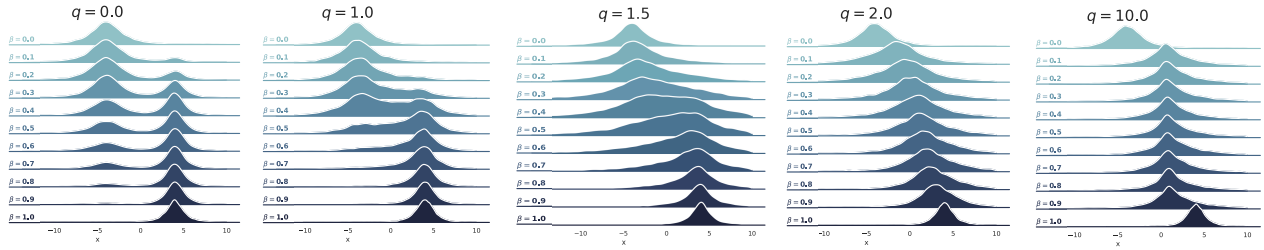


Figure 1: Intermediate densities between Student-t distributions, $t_{\nu=1}(-4, 3)$ and $t_{\nu=1}(4, 1)$ for various q-paths and 10 equally spaced β, Note that $\nu = 1$ corresponds to $q = 2$, so that the $q = 2$ path stays within the q-exponential family.

Annealing between Student-t Distributions In Fig. 2, we consider annealing between two 1-dimensional Student-t distributions. We set $q = 2$, which corresponds to $\nu = 1$ with $\nu = (3 - q)/(q - 1)$, and use the same mean and variance as the Gaussian example in Fig. 2, with $\pi_0(z) = t_{\nu=1}(-4, 3)$ and $\pi_1(z) = t_{\nu=1}(4, 1)$. For this special case of both endpoint distributions within a parametric family, we can ensure that the $q = 2$ path stays within the q-exponential family of Student-t distributions, just as the $q = 1$ path stayed within the Gaussian family in Fig. 2.

Comparing the $q = 0.5$ and $q = 0.9$ paths in the Gaussian case (Fig. 2) with the $q = 1.0$ and $q = 1.5$ path for the Student-t family with $q = 2$, we observe that mixing behavior appears to depend on the relation between the q-path parameter and the order of the q-exponential family of the endpoints. For our experiments in the main text, we did not find benefit to increasing $q > 1$. However, the toy example above indicates that $q > 1$ may be useful in some settings, for example involving heavier tailed distributions.

As $q \to \infty$, the power mean (15) approaches the $\min$ operation as $1 - q \to -\infty$. In the Gaussian case in Fig. 2, we see that, even at $q = 2$, intermediate densities for all β appear to concentrate in regions of low density under both π_0 and π_T. However, for the heavier-tailed Student-t distributions, we must raise the q-path parameter significantly to observe similar behavior.

D.3 MOMENT-AVERAGED PATH AS A GENERALIZED MEAN

While our q-paths can take arbitrary unnormalized density functions $\mathbf{u} = \big(\tilde{\pi}_0(z), \tilde{\pi}_1(z)\big)$ as input arguments for the generalized mean, we can reinterpret the moment-averaging path as a generalized mean over the natural parameters $\mathbf{u} = \big(\theta_0, \theta_1\big)$. We contrast the difficulty of inverting the function $h(\theta)$ for the moments path (which involves the Legendre transform), against the simple form of the geometric or q-paths as arithmetic means in the parameter space θ as in Appendix D.2.

The moment-averaged path is defined using a convex combination of the dual parameter vectors (Grosse et al., 2013), for the restricted case where $\pi_0(z)$ and $\pi_1(z)$ are members of the same exponential family, with parameters θ_0 and θ_1

$$\eta(\theta_\beta) = (1 - \beta)\, \eta(\theta_0) + \beta\, \eta(\theta_1)\,. \tag{49}$$

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

$$\theta_\beta = \eta^{-1}\big((1 - \beta)\, \eta(\theta_0) + \beta\, \eta(\theta_1)\big)\,. \tag{50}$$

Comparing to the form of Eq. (15), we can interpret the moment-averaging path as a generalized mean, with the natural parameters $\mathbf{u} = (\boldsymbol{\theta_0}, \boldsymbol{\theta_1})$ as inputs and the sufficient statistic function as the transformation $h(\theta) = \eta(\theta)$, although calculating the inverse is difficult in practice.

This observation highlights the convenience of working with generalized means in unnormalized density function space as in q-paths. When constructing paths from generalized means in parameter space θ, one may have to calculate normalization constants or consider the entire domain of the density function. By contrast, the expression for q-paths in Eq. (2) only involves inverting a scalar function at each point in the input sample space z.

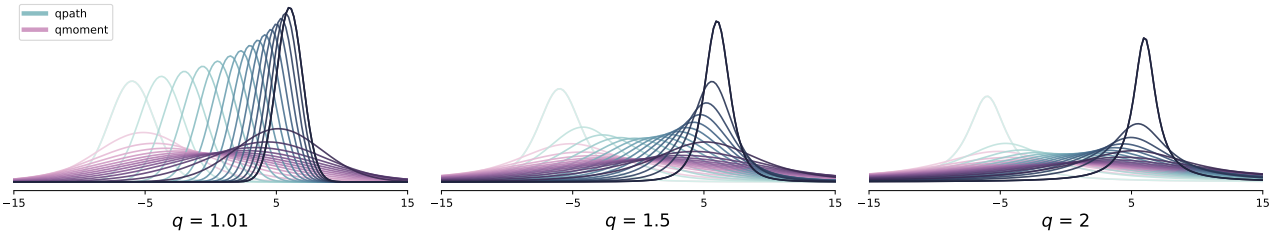


Figure 2: We visualize the escort-moments path for Student-t endpoints with $t_\nu(-4, 3)$ and $t_\nu(4, 1)$ for various $\nu = (3-q)/(1-q)$. We compare the corresponding q-path, whose intermediate densities remain within the q-exponential family, to the escort-moments path (Eq. (57)). Note, $q = 1.01$ closely resembles the moment-averaged path of Grosse et al. (2013).

D.4 ESCORT MOMENT-AVERAGED PATH

While exponential families are ubiquitous throughout machine learning, whether via common parametric distributions such as Gaussians or energy-based models such as (Restricted) Boltzmann Machines, models involving the q-exponential function have received comparatively little attention in machine learning. Nevertheless, we derive an analogue of the moment-averaged path for endpoint distributions within the same q-exponential family, with several parametric examples in App. D. We begin by recalling the definition,

$$\pi_{\theta,q}(z) = g(z)\, \exp_q \left\{\theta \cdot \phi_q(z) - \psi_q(\theta)\right\}. \tag{51}$$

where $g(z)$ indicates a base distribution and $\psi_q(\theta)$ denotes the q-free energy, which is convex as a function of the parameter θ (Amari and Ohara, 2011).

As in the case of the exponential family, differentiating the q-free energy yields a dual parameterization of the q-exponential family (Amari and Ohara, 2011). However, the standard expectation is now replaced with the *escort* expectation (Naudts, 2011)

$$\eta_q(\theta) = \nabla_\theta \psi_q(\theta) = \int \frac{\tilde{\pi}_\theta^{(q)}(z)^q}{\int \tilde{\pi}_\theta^{(q)}(z)^q} \cdot \phi(z) dz \tag{52}$$

$$:= \mathbb{E}_{\Pi_q(\theta)}[\phi(z)] \tag{53}$$

where $\Pi_q(\theta) \propto \tilde{\pi}_{\theta,q}(z)^q$ is the escort distribution for a given for $\tilde{\pi}_{\theta,q}$ in a parametric q-exponential family. This reduces to the standard expectation for $q = 1$ as in Eq. (9).

We propose the escort moment-averaging path for endpoints within a q-exponential family, using linear mixing in the dual parameters. Letting the function $\eta_\Pi(\theta)$ output the escort expected sufficient statistics for a q-exponential family distribution with parameter θ,

$$\eta_{\Pi_q}(\theta_\beta) = (1-\beta)\, \eta_{\Pi_q}(\theta_0) + \beta\, \eta_{\Pi_q}(\theta_1) \tag{54}$$

To provide a concrete example of the escort moment-averaging path in Fig. 2, we consider the Student-t distribution, which uses the same first- and second-order sufficient statistics as a Gaussian distribution and a degrees of freedom parameter ν that specifies the order of the q-exponential function for $q \geq 1$. This parameter induces heavier tails than a standard Gaussian, which appears as a special case as $q \to 1$ and $\exp_q(u) \to \exp(u)$.

In Fig. 2, we observe that the escort moments path spreads probability mass more widely than the q-path, which matches the observations of Grosse et al. (2013) in comparing the moment-averaging path to the geometric path for exponential family endpoints. Note that the q-path remains within the q-exponential family as shown in Appendix D.2.

We proceed to derive a closed form expression for the parameters of intermediate distributions along the escort moment-averaged path between Student-t endpoints.

D.5 ESCORT MOMENT-AVERAGED PATH WITH STUDENT-t ENDPOINTS

For the case of the Student-t distribution with degrees of freedom parameter ν , the escort distribution is *also* a Student-t distribution, but with $\nu' = \nu + 2$ and a rescaling of the covariance matrix $\frac{1}{Z_\Pi(\Sigma)} t_\nu(z; \mu, \Sigma)^q = t_{\nu+2}(z; \mu, \frac{\nu}{\nu+2}\Sigma)$ (Tanaka

2010, Matsuzoe 2017).

Finding the escort moment-averaged path thus becomes a moment matching problem over Student-t distributions with a different ν. We seek to find $\pi_\beta(z) = t_\nu(z; \mu_\beta, \Sigma_\beta)$ such that the expected sufficient statistics, under the escort distribution $\Pi_\beta(z) = t_{\nu+2}(z; \mu_\beta, \frac{\nu}{\nu+2}\Sigma_\beta)$, are equal to

$$\mathbb{E}_{\Pi_\beta}[z] = (1-\beta)\,\mathbb{E}_{\Pi_0}[z] + \beta\,\mathbb{E}_{\Pi_1}[z] \tag{55}$$

$$\mathbb{E}_{\Pi_\beta}\left[zz^T\right] = (1-\beta)\,\mathbb{E}_{\Pi_0}\left[zz^T\right] + \beta\,\mathbb{E}_{\Pi_1}\left[zz^T\right] \tag{56}$$

where optimization is over the parameters of the distribution $t_\nu(z; \mu_\beta, \Sigma_\beta)$. Note that $\mathbb{E}_{\Pi_\beta}[z] = \mu_\beta$ since the mean is unchanged for the escort distribution, whereas $\mathbb{E}_{\Pi_\beta}\left[zz^T\right] = \Sigma_{\Pi_\beta} + \mu_{\Pi_\beta}\mu_{\Pi_\beta}^T = \frac{\nu}{\nu+2}\Sigma_\beta + \mu_\beta\mu_\beta^T$.

Following similar derivations as in Grosse et al. (2013) Sec. 4 using the escort expressions, we have

$$\begin{aligned} \mu_\beta = \mu_{\Pi_\beta} &= (1-\beta)\mu_0 + \beta\mu_1 \\ \Sigma_\beta = \frac{\nu+2}{\nu}\Sigma_{\Pi_\beta} &= (1-\beta)\Sigma_0 + \beta\Sigma_1 + \frac{\nu+2}{\nu}\beta(1-\beta)(\mu_1-\mu_0)(\mu_1-\mu_0)^T \end{aligned} \tag{57}$$

which implies that the escort moment-averaged distribution has the form $t_\nu(z; \mu_\beta, \Sigma_\beta)$, with the same degrees of freedom ν as in the original q-exponential family.

E ADDITIONAL EXPERIMENTS FOR PARAMETRIC ENDPOINT DISTRIBUTIONS

In these experiments, we consider using Annealed Importance Sampling (AIS) to estimate the partition function ratio for well-separated 1-d Gaussian ($q = 1$) and Student-t ($q > 1$) endpoint distributions. Our goal is to compare the performance of the moment-averaging or escort-moment averaging paths, which are limited to the case of parametric endpoints distributions, with the more general q-paths.

Gaussian To compare q-paths against the moment-averaging path (Grosse et al., 2013), we anneal between $\pi_0 = \mathcal{N}(-4, 3)$ and $\pi_1 = \mathcal{N}(4, 1)$. Similarly, we anneal between $\pi_0 = t_{\nu=1}(-4, 3)$ and $\pi_1 = t_{\nu=1}(4, 1)$, where $\nu = 1$ corresponds to $q = 2$, to compare against the escort moment-averaged path in Appendix D. For all experiments, we use use parallel runs of Hamiltonian Monte Carlo (HMC) (Neal, 2011) to obtain 2.5k independent samples from $\tilde{\pi}_{\beta,q}(z)$ using K linearly spaced β_t between $\beta_0 = 0$ and $\beta_K = 1$. We perform a grid search over 20 log-spaced $\delta \in [10^{-5}, 10^{-1}]$ and report the best $q = 1 - \delta$.

Results are shown in Fig. 3, where we observe q-paths outperform the geometric path in both cases, as well as the moment and q-moments paths which have closed-form expressions and exact samples. In App. D.2, we provide additional analysis for annealing between two Student-t distributions.

Student-t Since the Student-t family generalizes the Gaussian distribution to $q \neq 1$, we can run a similar experiment annealing between two Student-t distributions. We set $q = 2$, which corresponds to $\nu = 1$ with $\nu = (3-q)/(q-1)$, and use the same mean and variance as the Gaussian example in Fig. 2 or Student-t example in Fig. 2 with $\pi_0(z) = t_{\nu=1}(-4, 3)$ and $\pi_1(z) = t_{\nu=1}(4, 1)$.

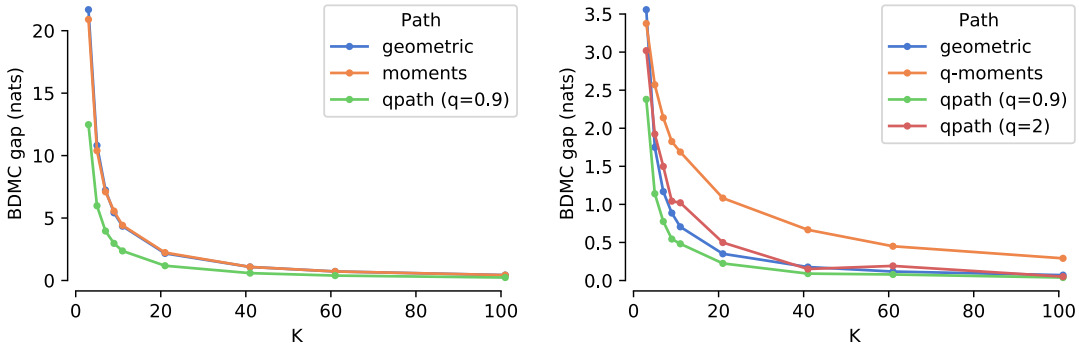
In Fig. 3, we compare the escort-moment averaging path with $q = 2$ to the geometric path and various q-paths. As shown in Appendix D.2, the q-path with $q = 2$ stays within the q-exponential family. The escort-moment averaging path does not outperform q-paths, which may be surprising since it appears to achieve interesting mass covering behavior in Fig. 2. As in the Gaussian case, we see that q-paths with $q \neq 2$ can achieve improvements even when the endpoints Student-t distributions use $q = 2$.

F SUM AND PRODUCT IDENTITIES FOR q-EXPONENTIALS

In this section, we prove two lemmas which are useful for manipulation expressions involving q-exponentials, for example in moving between Eq. (7) and Eq. (8) in either direction.

Lemma 1. *Sum identity*

$$\exp_q\left(\sum_{n=1}^N x_n\right) = \prod_{n=1}^N \exp_q\left(\frac{x_n}{1 + (1-q)\sum_{i=1}^{n-1} x_i}\right) \tag{58}$$



(a) $\mathcal{N}(-4,3) \to \mathcal{N}(4,1)$ (b) $t_{\nu=1}(-4,3) \to t_{\nu=1}(4,1)$

Figure 3: BDMC gaps for various paths on toy models. q-Paths out perform both the moments and the escort-moments path, both of which make use of parametric endpoint assumptions. Best q out of 20 shown.

Lemma 2. *Product identity*

$$\prod_{n=1}^{N} \exp_q(x_n) = \exp_q\left(\sum_{n=1}^{N} x_n \cdot \prod_{i=1}^{n-1}(1+(1-q)x_i)\right) \tag{59}$$

F.1 PROOF OF LEMMA 1

Proof. We prove by induction. The base case ($N=1$) is satisfied using the convention $\sum_{i=a}^{b} x_i = 0$ if $b < a$ so that the denominator on the RHS of Eq. (58) is 1. Assuming Eq. (58) holds for N,

$$\exp_q\left(\sum_{n=1}^{N+1} x_n\right) = \left[1+(1-q)\sum_{n=1}^{N+1} x_n\right]_+^{1/(1-q)} \tag{60}$$

$$= \left[1+(1-q)\left(\sum_{n=1}^{N} x_n\right)+(1-q)x_{N+1}\right]_+^{1/(1-q)} \tag{61}$$

$$= \left[\left(1+(1-q)\sum_{n=1}^{N} x_n\right)\left(1+(1-q)\frac{x_{N+1}}{1+(1-q)\sum_{n=1}^{N} x_n}\right)\right]_+^{1/(1-q)} \tag{62}$$

$$= \exp_q\left(\sum_{n=1}^{N} x_n\right)\exp_q\left(\frac{x_{N+1}}{1+(1-q)\sum_{n=1}^{N} x_n}\right) \tag{63}$$

$$= \prod_{n=1}^{N+1} \exp_q\left(\frac{x_n}{1+(1-q)\sum_{i=1}^{n-1} x_i}\right) \text{ (using the inductive hypothesis)} \tag{64}$$

□

F.2 PROOF OF LEMMA 2

Proof. We prove by induction. The base case ($N=1$) is satisfied using the convention $\prod_{i=a}^{b} x_i = 1$ if $b < a$. Assuming Eq. (59) holds for N, we will show the $N+1$ case. To simplify notation we define $y_N := \sum_{n=1}^{N} x_n \cdot \prod_{i=1}^{n-1}(1+(1-q)x_i)$.

Then,

$$\prod_{n=1}^{N+1} \exp_q(x_n) = \exp_q(x_1) \left(\prod_{n=2}^{N+1} \exp_q(x_n) \right) \tag{65}$$

$$= \exp_q(x_0) \left(\prod_{n=1}^{N} \exp_q(x_n) \right) \qquad \text{(reindex } n \to n-1\text{)}$$

$$= \exp_q(x_0) \exp_q(y_N) \qquad \text{(inductive hypothesis)}$$

$$= \Big[(1 + (1-q) \cdot x_0) \, (1 + (1-q) \cdot y_N) \Big]_+^{1/(1-q)} \tag{66}$$

$$= \Big[1 + (1-q) \cdot x_0 + \big(1 + (1-q) \cdot x_0\big)(1-q) \cdot y_N \Big]_+^{1/(1-q)} \tag{67}$$

$$= \Big[1 + (1-q) \Big(x_0 + \big(1 + (1-q) \cdot x_0\big) y_N \Big) \Big]_+^{1/(1-q)} \tag{68}$$

$$= \exp_q \left(x_0 + \big(1 + (1-q) \cdot x_0\big) y_N \right) \tag{69}$$

Next we use the definition of y_N and rearrange

$$= \exp_q \left(x_0 + \big(1 + (1-q) \cdot x_0\big) \left(x_1 + x_2(1 + (1-q) \cdot x_1) + ... + x_N \cdot \prod_{i=1}^{N-1} (1 + (1-q) \cdot x_i) \right) \right)$$

$$= \exp_q \left(\sum_{n=0}^{N} x_n \cdot \prod_{i=1}^{n-1} (1 + (1-q) x_i) \right). \tag{70}$$

Then reindexing $n \to n+1$ establishes

$$\prod_{n=1}^{N+1} \exp_q(x_n) = \exp_q \left(\sum_{n=1}^{N+1} x_n \cdot \prod_{i=1}^{n-1} (1 + (1-q) x_i) \right). \tag{71}$$

□

G EXPERIMENTAL DETAILS AND RESULTS

Algorithm 1 ESS Heuristic for Q-paths

Input: Set of log weights $\{\log w_i\}_{i=1}^S$, random restarts M, sample variance σ
Output: q, β which minimizes ESS criterion from Chopin and Papaspiliopoulos (2020).
Initialize $\delta_0 = \max_i |\log w_i|$ and $\mathcal{L}_{\text{best}} = \infty$
for j from 1 to M **do**
  Initialize $\beta_0 = 1$, $q_0 = 1 - \rho^{-1}$ with $\rho \sim \mathcal{N}(\rho_0, \sigma)$
  Solve $\beta^*, q^* = \text{argmin}_{\beta,q} \mathcal{L}(\beta_0, q_0)$ with $\mathcal{L}$ defined in Eq. (39) using coordinate descent.
  if $\mathcal{L}(\beta^*, q^*) < \mathcal{L}_{\text{best}}$ **then**
    Set $q_{\text{best}} \leftarrow q^*$, $\beta_{\text{best}} \leftarrow \beta^*$, $\mathcal{L}_{\text{best}} \leftarrow \mathcal{L}(\beta^*, q^*)$
  end if
end for
return $q_{\text{best}}, \beta_{\text{best}}$

G.1 SEQUENTIAL MONTE CARLO

We follow the experimental setup from Ch. 17.3 of Chopin and Papaspiliopoulos (2020) using the preprocessed Pima Indians diabetes ($N = 768, D = 9$) and Sonar datasets ($N = 208, D = 61$) available at `https://particles-sequential-monte-carlo-in-python.readthedocs.io/en/latest/datasets.html`. The model is specified as:

$$p(w_j) = \mathcal{N}(0, 5^2) \qquad p(y_i|x_i, w) = \text{Bern}(p_i = \text{sigmoid}(x_i^T w)) \tag{72}$$

$$p(\theta) = \prod_{j=1}^{D} p(w_j) \qquad p(\mathcal{D}, \theta) = p(\theta) \prod_{i=1}^{N} p(y_i|x_i, w). \tag{73}$$

In Algorithm 1 we use $M = 100$ restarts and compute ρ in $\log_{10}$ space with a sample variance $\sigma = 0.1$ (i.e $q = 1 - 10^{-\rho}$ for $\rho \sim \mathcal{N}(\log_{10}(\rho_0), 0.1)$). For coordinate descent we use the modified Powell algorithm available from the scipy python library.

G.2 EVALUATING GENERATIVE MODELS USING AIS

Table 1: Settings for training and evaluating a variational autoencoder (VAE) generative model trained with thermodynamic variational objective (TVO) on the Omniglot dataset.

Configuration	Value
training examples	24,345
simulated examples	2,500
real test examples	8,070
epochs	5000
number of importance samples	50
number of TVO partitions	100
TVO partition schedule	log uniform ($\beta_1 = 0.025$)
decoder	[50, 200, 200, 784]
encoder	[784, 200, 200, 50]
batch size	100
activation function	tanh

REFERENCES

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

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

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.

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

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.

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.

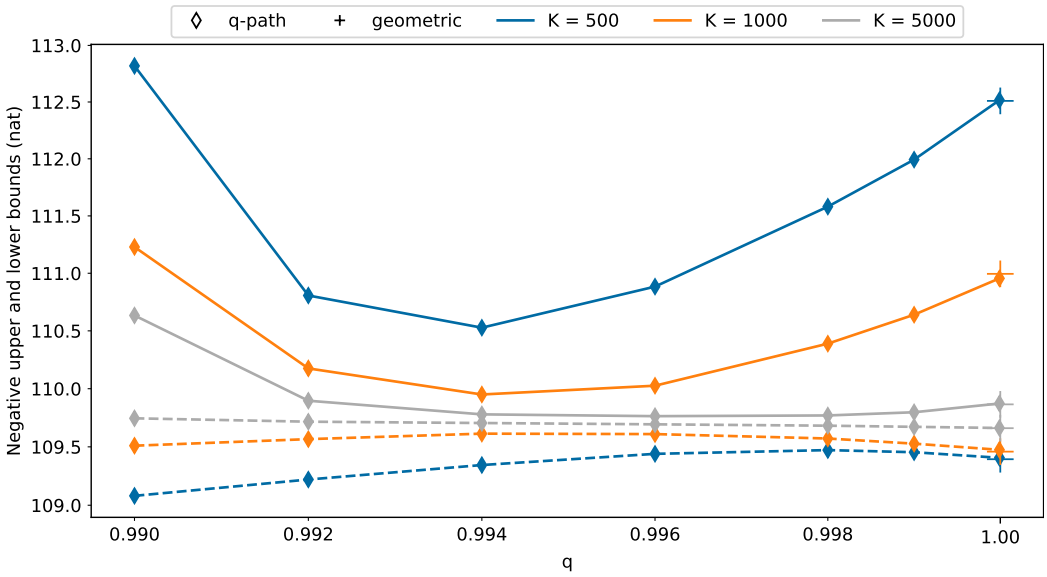


Figure 4: Stochastic lower and upper bounds produced by forward and reverse Hamiltonian AIS runs, for various numbers of annealing distributions (K) and q-values. Best viewed in colour.

Anders Meng. An introduction to variational calculus in machine learning. 2004.

Kevin P Murphy. Conjugate bayesian analysis of the gaussian distribution. *def*, 1(2σ2):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. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, page 113, 2011.

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. *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.