The Bayesian Learning Rule Beyond KL Geometry

Sophia Sklaviadis[*,1], **Wu Lin**[‡], **André F. T. Martins**[*],
Mário A. T. Figueiredo[*], **Thomas Möllenhoff**[†], **Mohammad Emtiyaz Khan**[†,§]

[*]Instituto de Telecomunicações and Instituto Superior Técnico, Universidade de Lisboa, Portugal
[†]RIKEN Center for Advanced Intelligence Project, Japan
[‡]University of Central Florida, USA
[§]TU Darmstadt and hessian.AI, Germany

[1]ssklaviadis@gmail.com

Abstract

The Bayesian learning rule (BLR) exploits the relation between Bayesian updating and the Kullback-Leibler divergence (KLD) to derive efficient additive updates for exponential-family variational posteriors. The KL geometry is powerful but restrictive. For instance, posteriors with parameter-dependent support can lead to undefined objectives. We ask which non-KL geometries preserve the additive BLR computation. Previous works that use f-divergences suggest that this is only possible with the KLD. Here, we show that this is not true: iteratively locally approximately conjugate updates can be derived for a broad class of mixed-space Bregman or Fenchel-Young discrepancies. These discrepancies not only generalize maximum-entropy families and dual coordinates, but also lead to additive updates including the classical BLR as a special case. This generalized BLR leads to new algorithms that *cannot* be derived from the KL-based BLR. For the Tsallis negative entropy, the generalized BLR yields bounded-support and heavy-tailed q-Gaussian posteriors and an easy Newton-like implementation. Experiments up to CIFAR-100 illustrate controllable support and tail behavior and competitive performance. Code is available at https://github.com/team-approx-bayes/qvon.

1 Introduction

Variational Bayes views Bayesian inference as optimization over variational posteriors that strike a good balance between fitting the data well and adhering to the prior. Khan and Rue [2023]'s Bayesian learning rule (BLR) algorithm extends classical Bayes' rule by using iteratively locally approximately conjugate, additive updates to search for optimal exponential-family variational posteriors. The updates stem from the interaction of the Shannon entropy or KL geometry and the exponential family which is computationally powerful but can also be restrictive: the BLR's computation is tied to fixed-support exponential families, and posterior families with variable support or adaptive tail behavior fall outside the scope of the KL-based BLR. Here, we address a structural question raised by the BLR: what generalizations of the variational objective preserve iteratively locally approximately conjugate, additive updates but free the BLR from its KL restriction?

Recently, Knoblauch et al. [2022] propose a generalization of variational Bayes which treats the variational family, the loss function, and the divergence between the posterior and prior as independent design choices. In contrast, our generalization preserves the BLR's couplings between prior and entropic regularizer, and between prior and posterior families. Alquier [2021] also proposes an additive mirror-descent update that is similar to the BLR, which optimizes the sum of an f-divergence and an expected loss. Only when the f-divergence coincides with the KLD is it possible to separate

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

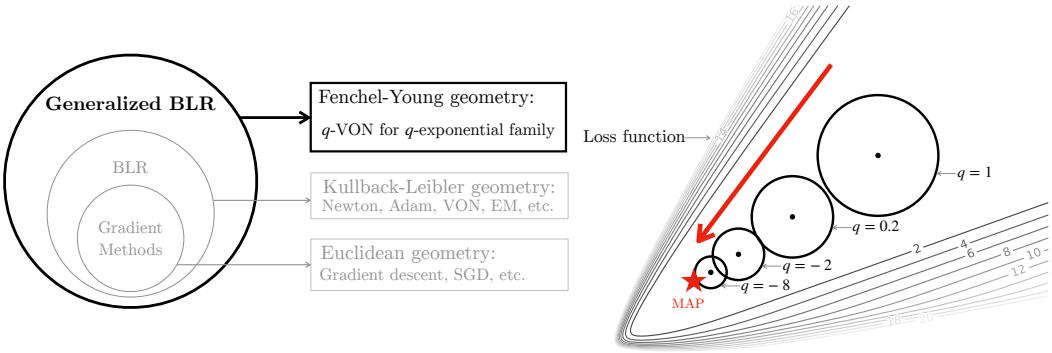


Figure 1: **Left:** We generalize the BLR by exploiting the geometry of Fenchel-Young discrepancies. We recover as special cases the BLR and the learning algorithms derived from it, and derive new algorithms using bounded-support and heavy-tailed q-exponential families. **Right:** Unlike Gaussians, bounded-support q-Gaussians allow control of the support by varying q. Thus, we can fit posteriors that closely track the MAP. The illustrated loss is that of 2D Bayesian logistic regression [Murphy, 2012]. Each circle indicates the first standard deviation, and outer loss contours indicate higher loss values. The Gaussian ($q = 1$) fit is farther from the MAP because its full support assigns non-negligible mass to high-loss regions; bounded-support q-Gaussians truncate this exploration.

the f-divergence ratio into a difference of linear expectations.[1] This separation gives the relevant instances of the BLR's additive updates their locally approximately conjugate interpretation: prior information and a local approximation of the expected loss are combined in terms of the canonical parameters of the posterior family. We preserve this connection between additive updates and conjugacy beyond the KLD.

In this paper, we use mixed-space Bregman or Fenchel-Young discrepancies that recover the KLD and BLR, but encompass a broad class of discrepancies generated by convex functions, including those associated with generalized entropies and proper scoring rules [Dawid, 2007]. We optimize a generalized variational objective using dual coordinates of a relaxed class of reference potentials in place of normalized priors, and generalized Fenchel posteriors in place of Gibbs posteriors. The resulting information-geometric algorithm has the same additive computation as the BLR, and admits a dual-space analogue of Bissiri et al. [2016]'s sequential coherence, in contrast to updates derived from f-divergence-regularized objectives. By KL or mixed-space Bregman geometry we mean the convex-analytic structure defined by a reference potential and convex function, which induce a mixed Bregman divergence: the potential family and convex function determine the maximum entropy posterior family, dual coordinates, and a Hessian metric. Our terminology is motivated by the fact that a mixed Bregman divergence is the slack in a Fenchel-Young inequality.[2]

Our contributions are: (1) We identify mixed-space Bregman or Fenchel-Young discrepancies, as defined by Amari and Cichocki [2010], Martins et al. [2022] but only first connected to variational Bayes in Sklaviadis et al. [2025], as a class that preserves BLR-style additive canonical updates. (2) We propose a generalized BLR update beyond the KL geometry and show that the classical BLR is recovered as a special case (Sec. 3). (3) We instantiate the generalized BLR update for the Tsallis negative entropy, yielding parameter-dependent bounded-support and heavy-tailed q-Gaussian Newton-like variational learning methods (Sec. 4, App. F). Our new Newton-like optimizers require changing a few lines of code in state-of-the-art Improved Variational Online Newton (IVON) [Shen et al., 2024], a scalable Gaussian BLR-style optimizer for deep networks, to obtain heavy-tailed or bounded-support noise injection, maintain non-Gaussian posteriors, and target a generalized Bayesian objective using the maximum entropy (maxent) geometry associated with each objective. These optimizers enable generalized Bayesian learning beyond the KLD at scale (Sec. 5, Apps. G and H).

[1] Uniquely, the KLD belongs both to the classes of f- and Bregman divergences [Amari, 2009].

[2] We use *Fenchel* rather than *Legendre* deliberately. In convex analysis, the Legendre transform is the differentiable special case of the Fenchel conjugate [Boyd and Vandenberghe, 2004, Touchette, 2005].

Parameter-dependent bounded-support q-Gaussian posteriors allow the posterior mean to closely track MAP solutions (Figs. 1 and 2). By contrast, Gaussian variational means are far enough away from the MAP to avoid sampling parameter-space regions with high-loss. The heavy-tailed q-Gaussian posterior updates we derive are a different parametrization of Student-t distributions, which are studied in robust Bayesian inference. Various aspects of robust Bayes can be reinterpreted in our framework, with the benefit of easy conjugate priors, but drawing these connections is beyond the scope of this paper; as is demonstrating state-of-the-art at scale. We show experiments up to CIFAR-100 (App. G), and defer large experiments, for example on quantization-aware training, pruning, and language model training with bounded variance to future work. We develop the theoretical framework and check that the proposed algorithms for variational training with controllable support and tail behaviors are competitive with their Gaussian IVON counterpart [Khan et al., 2018, Osawa et al., 2019, Shen et al., 2024], without additional tuning.

2 KLD Geometry and the BLR

The BLR relies on two levels of a common convex duality. At the density-space level, the Gibbs-Donsker-Varadhan variational formula is the Fenchel duality between the log-Laplace functional of a tilted prior, and the relative entropy with respect to that prior. Equivalently, if the prior is represented by a reference potential η_0, the KLD is the Fenchel-Young slack generated by the Shannon negentropy and that potential. The density-space duality, restricted by finite moment constraints, results in the maximum entropy characterization of parametric exponential families: with constant base measure, the exponential family $\pi_{\boldsymbol{\lambda}}$ with canonical parameter $\boldsymbol{\lambda}$ maximizes the Shannon entropy among all densities with the same moments, and with a non-constant base measure it maximizes the relative entropy with respect to that base measure. At the parametric level, restricting the potential to the finite-dimensional linear class $\eta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \langle \boldsymbol{\lambda}, \mathbf{T}(\boldsymbol{\theta}) \rangle$ induces the cumulant $A(\boldsymbol{\lambda})$. Under regularity and minimality, $\boldsymbol{\mu} = \nabla A(\boldsymbol{\lambda})$, and $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ are the dual coordinates on which the BLR update is based.

Fix a measurable space $(\Theta, \mathcal{A})$ and a σ-finite reference measure h. Let $\mathcal{P}$ be a convex class of probability densities with respect to h. Potentials η belong to a vector space of measurable functions in duality with a linear space containing $\mathcal{P}$, with pairing $\mathbb{E}_\pi[\eta] = \int_\Theta \eta(\boldsymbol{\theta})\pi(\boldsymbol{\theta})dh(\boldsymbol{\theta})$. For the Shannon negentropy $\Omega_{\mathrm{Sh}} = -\mathcal{H}(\pi) = \int_\Theta \pi(\boldsymbol{\theta}) \log \pi(\boldsymbol{\theta}) dh(\boldsymbol{\theta})$, the Fenchel conjugate over densities

$$\Omega^*_{\mathrm{Sh}}(\eta) = \sup_{\pi \in \mathcal{P}} \{\mathbb{E}_\pi[\eta(\boldsymbol{\theta})] - \Omega_{\mathrm{Sh}}(\pi)\} = \log \int_\Theta e^{\eta(\boldsymbol{\theta})} dh(\boldsymbol{\theta}),$$

is log-normalization, whenever the integral is finite. The corresponding Fenchel-Young slack is

$$\mathbb{D}_{-\mathcal{H}}(\eta, \pi) = \Omega^*_{\mathrm{Sh}}(\eta) + \Omega_{\mathrm{Sh}}(\pi) - \mathbb{E}_\pi[\eta(\boldsymbol{\theta})] \geq 0.$$

The KLD can be derived as the Fenchel-Young slack generated by Shannon negentropy and a reference potential. In particular, if a prior density is represented by potential η_0, as $\pi_0(\boldsymbol{\theta}) = \exp\left(\eta_0(\boldsymbol{\theta}) - \Omega^*_{\mathrm{Sh}}(\eta_0)\right)$, then $\log \pi_0(\boldsymbol{\theta}) = \eta_0(\boldsymbol{\theta}) - \Omega^*_{\mathrm{Sh}}(\eta_0)$. Hence, for any density π,

$$\begin{aligned}
\mathbb{D}_{-\mathcal{H}}(\eta_0, \pi) &= \Omega^*_{\mathrm{Sh}}(\eta_0) + \int_\Theta \pi(\boldsymbol{\theta}) \log \pi(\boldsymbol{\theta}) dh(\boldsymbol{\theta}) - \mathbb{E}_\pi[\eta_0(\boldsymbol{\theta})] \\
&= \int_\Theta \pi(\boldsymbol{\theta}) \left[\log \pi(\boldsymbol{\theta}) - \eta_0(\boldsymbol{\theta}) + \Omega^*_{\mathrm{Sh}}(\eta_0)\right] dh(\boldsymbol{\theta}) = \mathbb{D}_{\mathrm{KL}}(\pi \,\|\, \pi_0). \qquad (1)
\end{aligned}$$

Additive constants in η_0 do not matter, since they change $\Omega^*_{\mathrm{Sh}}(\eta_0)$ by the same constant.

The Gibbs or Donsker-Varadhan variational formula follows by applying the same slack to the tilted potential $\eta_0 - \ell$. The normalizer of the tilted prior is

$$\mathcal{Z} = \int_\Theta e^{-\ell(\boldsymbol{\theta})} \pi_0(\boldsymbol{\theta}) dh(\boldsymbol{\theta}) = \exp\left(\Omega^*_{\mathrm{Sh}}(\eta_0 - \ell) - \Omega^*_{\mathrm{Sh}}(\eta_0)\right).$$

Therefore, for every $\pi \in \mathcal{P}$ with $\pi \ll \pi_0$,

$$\begin{aligned}
\mathbb{E}_\pi[\ell(\boldsymbol{\theta})] + \mathbb{D}_{\mathrm{KL}}(\pi \,\|\, \pi_0) &= \mathbb{E}_\pi[\ell(\boldsymbol{\theta})] + \Omega_{\mathrm{Sh}}(\pi) - \mathbb{E}_\pi[\eta_0(\boldsymbol{\theta})] + \Omega^*_{\mathrm{Sh}}(\eta_0) \\
&= -\log \mathcal{Z} + \Omega^*_{\mathrm{Sh}}(\eta_0 - \ell) + \Omega_{\mathrm{Sh}}(\pi) - \mathbb{E}_\pi[\eta_0(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta})] \\
&= -\log \mathcal{Z} + \mathbb{D}_{-\mathcal{H}}(\eta_0 - \ell, \pi) \geq -\log \mathcal{Z}. \qquad (2)
\end{aligned}$$

Taking the infimum over candidates π we get the Gibbs-Donsker-Varadhan variational formula

$$-\log \mathcal{Z} = \inf_{\pi \in \mathcal{P}: \pi \ll \pi_0} \mathbb{E}_\pi[\ell(\boldsymbol{\theta})] + \mathbb{D}_{\mathrm{KL}}(\pi \,\|\, \pi_0). \tag{3}$$

Conversely, sharpness of the Gibbs bound uniquely characterizes the KL regularizer (App. A). Equality in Eq. 3 holds when the Fenchel-Young slack in Eq. 2 vanishes, that is, when π is the normalized density induced by the tilted potential $\eta_0 - \ell$, namely the Gibbs posterior:

$$\pi_*(\boldsymbol{\theta}) = \exp\left(\eta_0(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}) - \Omega^*_{\mathrm{Sh}}(\eta_0 - \ell)\right) = \frac{e^{-\ell(\boldsymbol{\theta})}\pi_0(\boldsymbol{\theta})}{\mathcal{Z}}.$$

This duality appears in several literatures. In statistical mechanics and information theory, free-energy minimization and maximum Shannon entropy subject to moment constraints yield exponential families [Gibbs, 1902, Jaynes, 1957, Williams, 1980, Zellner, 1988]. In generalized Bayesian inference, the negative log-likelihood is replaced by a general loss while retaining the KLD penalty to the prior [Zhu et al., 2014, Bissiri et al., 2016, Ghosh and Basu, 2016, Jewson et al., 2018, Miller and Dunson, 2019]. In PAC-Bayesian learning theory, the KLD measures the complexity of moving from a fixed prior to a data-dependent posterior [Zhang, 1999, Catoni, 2007, Grünwald and Mehta, 2020]. In variational inference, the ELBO is equivalently a KLD-regularized expected-fit objective or a KL projection of the exact posterior onto a tractable family [Wainwright et al., 2008, Hoffman et al., 2013, Blei et al., 2017]. Instead of replacing the KLD in Eq. 3, we use the Fenchel-Young structure that generates the KLD as the template for the mixed-space Bregman construction in Sec. 3.[3]

2.1 Exponential Family Posteriors from Moment-Constrained Maxent

When the posterior belongs to a regular, minimal exponential family, through a constant base measure determined by finite moment constraints, the duality between log-normalization and negative entropy characterizes the posterior as the unique Shannon maxent distribution. To see this, first we fix a base measure $b(\boldsymbol{\theta})d\boldsymbol{\theta}$ and a finite sufficient statistic

$$\mathbf{T}(\boldsymbol{\theta}) = (T_1(\boldsymbol{\theta}), ..., T_P(\boldsymbol{\theta}))\,. \tag{4}$$

For each $\boldsymbol{\lambda} \in \mathbb{R}^P$, the linear potential $\eta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \langle \boldsymbol{\lambda}, \mathbf{T}(\boldsymbol{\theta})\rangle$ defines an exponential tilt $e^{\eta_{\boldsymbol{\lambda}}(\boldsymbol{\theta})}b(\boldsymbol{\theta})d\boldsymbol{\theta}$, which is not, in general, a probability measure. For $\boldsymbol{\lambda} \in \Lambda = \{\boldsymbol{\lambda} \in \mathbb{R}^P : \int b(\boldsymbol{\theta}) \exp\left(\eta_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\right) d\boldsymbol{\theta} < \infty\}$, the unique additive correction is the cumulant or log-partition function $A(\boldsymbol{\lambda}) = \log \int b(\boldsymbol{\theta}) \exp\left(\langle \boldsymbol{\lambda}, \mathbf{T}(\boldsymbol{\theta})\rangle\right) d\boldsymbol{\theta}$, defined on Λ, and the density induced by h and $\mathbf{T}$ has the familiar exponential family form

$$\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = b(\boldsymbol{\theta}) \exp\left(\langle \boldsymbol{\lambda}, \mathbf{T}(\boldsymbol{\theta})\rangle - A(\boldsymbol{\lambda})\right). \tag{5}$$

Note that $\boldsymbol{\lambda}$ is introduced as the coefficient of the finite-dimensional potential, and $A(\boldsymbol{\lambda})$ is derived as the log-normalizer of the corresponding exponential tilt, which is the measure-theoretic operation that uses the potential $\eta_{\boldsymbol{\lambda}}$ to tilt the base measure, thereby characterizing the exponential family form.[4]

Differentiating $\int \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})d\boldsymbol{\theta} = 1$ with respect to $\boldsymbol{\lambda}$ we obtain the moment map $\nabla_{\boldsymbol{\lambda}} A(\boldsymbol{\lambda}) = \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\mathbf{T}(\boldsymbol{\theta})] = \boldsymbol{\mu}(\boldsymbol{\lambda})$. Thus, natural and expectation parameters form a Legendre dual coordinate system. For simplicity, we assume the base measure is constant. In this case, the member of the exponential family with moments $\boldsymbol{\mu}(\boldsymbol{\lambda})$ is the unique Shannon entropy maximizer among all densities $\pi \in \mathcal{P}$, that satisfy the same moment constraints (see App. B.1 for a proof).[5]

2.2 The BLR in Exponential Family Coordinates

Khan and Rue [2023] apply Eq. 3 to unify various learning algorithms under the Bayesian Learning Rule (BLR). The BLR is obtained by restricting the DV optimization in Eq. 3 to the set of regular, minimal exponential families $\mathcal{Q} \subseteq \mathcal{P}$, and exploiting the $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ dual pair. Assuming, in addition, that the

[3] A different line of thought in generalized VI optimizes over a variational family by minimizing alternative divergences to the Bayesian posterior [Minka et al., 2005, Hernandez-Lobato et al., 2016, Li and Turner, 2016].

[4] The family is called regular if Λ is a nonempty open subset of $\mathbb{R}^P$. It is called minimal if there do not exist $(a, \mathbf{b}) \in \mathbb{R} \times \mathbb{R}^P$, $(a, \mathbf{b}) \neq (0, \mathbf{0})$, such that $a + \langle \mathbf{b}, \mathbf{T}(\boldsymbol{\theta})\rangle = 0$ for almost every $\boldsymbol{\theta}$. Under minimality, the natural parameter is identifiable, that is $\pi_{\boldsymbol{\lambda}_1} = \pi_{\boldsymbol{\lambda}_2}$ a.e. $\Rightarrow \boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_2$, and A is strictly convex.

[5] More generally, if the base measure is not constant, a corresponding statement can be made in terms of maximum relative Shannon entropy with respect to that base measure.

prior belongs to the same family, $\pi_0(\boldsymbol{\theta}) = b(\boldsymbol{\theta}) \exp\left(\langle \boldsymbol{\lambda}_0, \mathbf{T}(\boldsymbol{\theta})\rangle - A(\boldsymbol{\lambda}_0)\right)$, and taking b to be constant for simplicity, $\mathbb{D}_{-\mathcal{H}}(\log \pi_0, \pi_{\boldsymbol{\lambda}}) = A(\boldsymbol{\lambda}_0) + A^*(\boldsymbol{\mu}(\boldsymbol{\lambda})) - \langle \boldsymbol{\lambda}_0, \boldsymbol{\mu}\rangle$, where A^* is the Legendre dual of A, and coincides with the negative Shannon entropy $A^*(\boldsymbol{\mu}(\boldsymbol{\lambda})) = -\mathcal{H}(\pi_{\boldsymbol{\lambda}})$. Differentiating $\mathbb{D}_{-\mathcal{H}}(\log \pi_0, \pi_{\boldsymbol{\lambda}})$ with respect to the expectation parameter, $\nabla_{\boldsymbol{\mu}} \mathbb{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\lambda}} \,\|\, \pi_0) = \boldsymbol{\lambda} - \boldsymbol{\lambda}_0$.

For exponential families, the Fisher information matrix $\mathbf{F}(\boldsymbol{\lambda})$ can be shown to equal $\nabla^2_{\boldsymbol{\lambda}} A(\boldsymbol{\lambda})$, which allows us to use the chain rule to express natural gradients $\widetilde{\nabla}$ as reparametrized regular gradients, $\widetilde{\nabla}_{\boldsymbol{\lambda}} = \mathbf{F}(\boldsymbol{\lambda})^{-1} \nabla_{\boldsymbol{\lambda}} = \nabla_{\boldsymbol{\mu}}$. The exponential family restriction renders the natural gradient of the KLD affine in $\boldsymbol{\lambda}$, and the natural-gradient step into a mirror-descent step with an additive $\boldsymbol{\lambda}$ update

$$\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} - \rho \nabla_{\boldsymbol{\mu}} \left(\mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\ell(\boldsymbol{\theta})] + \mathbb{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\lambda}} \,\|\, \pi_0)\right) = (1-\rho)\boldsymbol{\lambda} - \rho \left(\nabla_{\boldsymbol{\mu}} \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\ell(\boldsymbol{\theta})] - \boldsymbol{\lambda}_0\right), \tag{6}$$

where ρ is a step size or learning rate.

Under conjugacy, we can simplify further. There exists a natural parameter $\boldsymbol{\lambda}'$ such that

$$\pi_{\boldsymbol{\lambda}'}(\boldsymbol{\theta}) = \frac{1}{\mathcal{Z}} \exp(-\ell(\boldsymbol{\theta}))\pi_0(\boldsymbol{\theta}) \Rightarrow \ell(\boldsymbol{\theta}) = \log \pi_0(\boldsymbol{\theta}) - \log \pi_{\boldsymbol{\lambda}'}(\boldsymbol{\theta}) - \log \mathcal{Z} \tag{7}$$

$$= \langle \boldsymbol{\lambda}_0 - \boldsymbol{\lambda}', \mathbf{T}(\boldsymbol{\theta})\rangle + A(\boldsymbol{\lambda}') - A(\boldsymbol{\lambda}_0) - \log \mathcal{Z} \tag{8}$$

Thus, $\mathbb{E}_{\pi_{\boldsymbol{\lambda}}} \ell(\boldsymbol{\theta}) = \langle \boldsymbol{\lambda}_0 - \boldsymbol{\lambda}', \boldsymbol{\mu}(\boldsymbol{\lambda})\rangle + \text{const.}$, and $\nabla_{\boldsymbol{\mu}} \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\ell(\boldsymbol{\theta})] = \boldsymbol{\lambda}_0 - \boldsymbol{\lambda}'$, so Eq. 6 simplifies to [6]

$$\boldsymbol{\lambda} \leftarrow (1-\rho)\boldsymbol{\lambda} + \rho\boldsymbol{\lambda}'.$$

For conjugate exponential family models, a single BLR step coincides with classical Bayes' rule. More generally, outside conjugacy, the BLR optimizes the Bayesian objective whose minimizer over $\mathcal{P}$ is the exact posterior. Many algorithms from optimization, statistics, and deep learning arise as special cases by varying the form of the exponential family. New algorithms have also been inspired by the general form of the BLR, including the Variational Online Newton algorithm that we extend in Sec. 4 and App. F to exemplify our proposed generalized BLR.

2.3 Limitations of the KLD and the BLR

KLD geometry has certain limitations. The KLD requires absolute continuity: if $\pi_{\boldsymbol{\lambda}}$ places mass outside the support of π_0, then $\mathbb{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\lambda}} \,\|\, \pi_0) = +\infty$. Thus posterior families with bounded, parameter-dependent support, and more generally families that cannot be represented as ordinary exponential families with fixed support, fall outside the scope of the KL-based BLR. Figs. 1 and 2 illustrate why bounded-support families are useful: they can avoid sampling high-loss regions and allow the variational mean to approach MAP solutions more closely than Gaussian approximations.

How then can we generalize the KLD so as to preserve the BLR's efficient, additive updating? In Sec. 3 we show that Fenchel-Young discrepancies mediate exactly such a generalization. They yield generalized maxent families with well-defined dual parameters and generalized canonical gradients, allowing BLR-style updates under bounded, parameter-dependent support and beyond. Within the class of updates obtained by minimizing the sum of expected cumulative loss and an f-divergence, sequential coherence as defined by Bissiri et al. [2016] singles out the Kullback-Leibler divergence. In that sense, an f-divergence generalization of the KLD, like the f-divergence-regularized updates of Alquier [2021], cannot be used to define Bayes-like additive updates. Our mixed-space Bregman generalization of the KLD avoids this problem. The generalized BLR respects a dual-space analogue of sequential coherence: losses additively tilt the reference potential, so processing losses sequentially or jointly yields the same tilted potential. An interesting question that we hope to address in the future is whether mixed-space Bregman divergences uniquely recover BLR-style updates when we also wish to preserve the linear expectation of the loss term in Eq. 3.

3 Fenchel-Young BLR

To derive a generalized BLR beyond KL geometry, first we introduce mixed-space Bregman divergences to which we refer interchangeably as Fenchel-Young discrepancies (FYD). We call $\mathbb{D}_\Omega(\eta, \pi)$

[6] Writing Eq. 6 in terms of densities by substituting the update for $\boldsymbol{\lambda}$ into the exponential family form of $\pi_{\boldsymbol{\lambda}}$ results in a multiplicative BLR update, which under conjugacy simplifies to

$$\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) \leftarrow \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})^{1-\rho} \pi_{\boldsymbol{\lambda}'}(\boldsymbol{\theta})^{\rho} = \frac{1}{\mathcal{Z}_\rho} \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})^{1-\rho} \left[\exp(-\ell(\boldsymbol{\theta}))\pi_0(\boldsymbol{\theta})\right]^{\rho}. \tag{9}$$

a *mixed-space* Bregman because its two arguments belong to different spaces: η is a dual potential, while π is a density. This generalization of the KLD forms the basis of our proposal and allows our construction to remain meaningful for generalized entropies that may be nonsmooth, extended-valued, or have parameter-dependent support. We show how FYDs lead to generalized maxent families, for which we can define dual coordinates. Finally, we derive a generalized variational objective and learning rule. FYDs are also studied as discriminative losses by Blondel et al. [2019], Martins et al. [2022], and their information-geometric properties are discussed in Amari and Cichocki [2010], Nielsen [2022]. Here, we use them in a novel way to generalize the Bayesian learning rule optimizing the generalized objective defined in Sklaviadis et al. [2025].

3.1 Generalizing the KLD to the Fenchel-Young Discrepancy

Fix a proper convex function $\Omega : \mathcal{P} \to \mathbb{R} \cup \{\infty\}$ on densities. The Fenchel conjugate of Ω

$$\Omega^*(\eta) = \sup_{\pi \in \mathcal{P}} \mathbb{E}_\pi[\eta(\boldsymbol{\theta})] - \Omega(\pi)$$

acts on measurable potentials $\eta : \Theta \to \mathbb{R} \cup \{\pm\infty\}$, for which the pairing $\langle \eta, \pi \rangle = \mathbb{E}_\pi[\eta(\boldsymbol{\theta})] = \int_\Theta \eta(\boldsymbol{\theta})\pi(\boldsymbol{\theta})dh(\boldsymbol{\theta})$ is well defined. For any $\eta \in \operatorname{dom}\Omega^*$ and $\pi \in \operatorname{dom}\Omega$ with finite pairing, the Fenchel-Young discrepancy is [Martins et al., 2022, Amari and Cichocki, 2010]

$$\mathbb{D}_\Omega(\eta, \pi) = \Omega^*(\eta) + \Omega(\pi) - \mathbb{E}_\pi[\eta(\boldsymbol{\theta})]. \tag{10}$$

The first argument is a reference potential, not necessarily the log-density of a normalized prior. Since every $\pi \in \mathcal{P}$ integrates to one, $\Omega^*(\eta + c) = \sup_{\pi \in \mathcal{P}} \{\mathbb{E}_\pi[\eta(\boldsymbol{\theta})] + c - \Omega(\pi)\} = \Omega^*(\eta) + c$, and therefore $\mathbb{D}_\Omega(\eta + c, \pi) = \mathbb{D}_\Omega(\eta, \pi)$. For general Ω, when the supremum defining $\Omega^*(\eta)$ is attained, any maximizer $\pi_\eta \in \arg\max_{\pi \in \mathcal{P}} \{\mathbb{E}_\pi[\eta(\boldsymbol{\theta})] - \Omega(\pi)\}$ is a density induced by the potential η. Equivalently, $\pi_\eta \in \partial\Omega^*(\eta)$ with respect to the (η, π) pairing.

3.2 Generalizing the Exponential Family

Dual coordinates related to the FYD can be defined for a much larger class of distributions than the exponential family, including deformed exponential families [Amari, 2016, Naudts, 2002]. Our generalization preserves the same logic we followed in Sec. 2.1 to derive exponential families and characterize their dual coordinates. We begin with a finite-dimensional $\eta_{\boldsymbol{\lambda}}$, thereby identify a convex potential $\Psi^*(\boldsymbol{\lambda}) = \Omega^*(\eta_{\boldsymbol{\lambda}})$, and only when the supremum defining $\Omega^*(\eta_{\boldsymbol{\lambda}})$ is attained by a density, do we recover the corresponding member $\pi_{\boldsymbol{\lambda}}$ of the induced maximum-entropy family. The generalized Fisher we use is distinct from both the invariant and conformally flat geometries of Amari et al. [2012], and it is a generalization of the metric associated with the derivation of deformed exponential families subject to linear (rather than escort) expectation constraints discussed in Naudts [2011].

Starting with a fixed measurable statistic $\mathbf{T}(\boldsymbol{\theta})$ as in Eq. 4, define the moment vector $\boldsymbol{\mu} = \mathbb{E}_\pi[\mathbf{T}(\boldsymbol{\theta})]$, and associate with it the Lagrange multiplier $\boldsymbol{\lambda} \in \mathbb{R}^P$. The corresponding finite-dimensional potential family is $\eta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \langle \boldsymbol{\lambda}, \mathbf{T}(\boldsymbol{\theta}) \rangle$, defined up to an additive constant.[7] We define Ψ to assign to each moment vector $\boldsymbol{\mu}$ the optimal value of the corresponding constrained minimization problem

$$\Psi(\boldsymbol{\mu}) = \inf_{\pi \in \mathcal{P}} \Omega(\pi) \text{ s.t. } \mathbb{E}_\pi[\mathbf{T}(\boldsymbol{\theta})] = \boldsymbol{\mu}, \text{ with domain } \mathcal{M} = \left\{ \mathbb{E}_\pi[\mathbf{T}(\boldsymbol{\theta})] : \pi \in \mathcal{P}, \Omega(\pi) < \infty \right\}. \tag{11}$$

We do not assume that the infimum in Eq. 11 is attained for every $\boldsymbol{\mu}$. When $\Omega = -\mathcal{H}$, Eq. 11 reduces to the classical Shannon maxent, and under the constant base measure convention of Sec. 2.1, whenever the minimizer exists it is the exponential family in Eq. 5 with natural parameter $\boldsymbol{\lambda} = \nabla A^*(\boldsymbol{\mu})$. In that case, $\Psi(\boldsymbol{\mu}) = A^*(\mu)$ and $\Psi^*(\boldsymbol{\lambda}) = A(\boldsymbol{\lambda})$.

Proposition 1. *For every $\boldsymbol{\lambda} \in \mathbb{R}^P$, $\Psi^*(\boldsymbol{\lambda}) = \Omega^*(\eta_{\boldsymbol{\lambda}})$.*

Proof. By definition of the convex conjugate,

$$\begin{aligned}\Psi^*(\boldsymbol{\lambda}) &= \sup_{\boldsymbol{\mu} \in \mathcal{M}} \{\langle \boldsymbol{\lambda}, \boldsymbol{\mu} \rangle - \Psi(\boldsymbol{\mu})\} = \sup_{\boldsymbol{\mu} \in \mathcal{M}} \sup_{\pi \in \mathcal{P}: \mathbb{E}_\pi[\mathbf{T}] = \boldsymbol{\mu}} \{\langle \boldsymbol{\lambda}, \boldsymbol{\mu} \rangle - \Omega(\pi)\} \\ &= \sup_{\pi \in \mathcal{P}} \left\{ \left\langle \boldsymbol{\lambda}, \mathbb{E}_\pi[\mathbf{T}(\boldsymbol{\theta})] \right\rangle - \Omega(\pi) \right\} = \sup_{\pi \in \mathcal{P}} \left\{ \mathbb{E}_\pi[\eta_{\boldsymbol{\lambda}}(\boldsymbol{\theta})] - \Omega(\pi) \right\} = \Omega^*(\eta_{\boldsymbol{\lambda}}). \quad \square\end{aligned}$$

[7] An additional constant statistic $T_0 = 1$ in $\mathbf{T}$ can absorb normalization constants into $\boldsymbol{\lambda}$.

Prop. 1 identifies the finite-dimensional convex potential associated with the potential family. However, it does not yet identify a family of densities, it only shows how the infinite dimensional moment problem projects to canonical coordinates. Next, we determine when a density obtained from a finite dimensional potential solves the infinite dimensional problem in Eq. 11.

Proposition 2. *Assume that for $\boldsymbol{\lambda} \in \mathbb{R}^P$ the supremum defining $\Omega^*(\eta_{\boldsymbol{\lambda}})$ is attained by $\pi_{\boldsymbol{\lambda}} \in \mathcal{P}$: $\Omega^*(\eta_{\boldsymbol{\lambda}}) = \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\eta_{\boldsymbol{\lambda}}(\boldsymbol{\theta})] - \Omega(\pi_{\boldsymbol{\lambda}})$. Define $\boldsymbol{\mu}(\boldsymbol{\lambda}) = \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\mathbf{T}(\boldsymbol{\theta})]$. Then $\pi_{\boldsymbol{\lambda}}$ minimizes Eq. 11.*

Proof. By the optimality of $\pi_{\boldsymbol{\lambda}}$ in the definition of $\Omega^*(\eta_{\boldsymbol{\lambda}})$, for every $\pi \in \mathcal{P}$, $\Omega(\pi) - \mathbb{E}_{\pi}[\eta_{\boldsymbol{\lambda}}(\boldsymbol{\theta})] \geq \Omega(\pi_{\boldsymbol{\lambda}}) - \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\eta_{\boldsymbol{\lambda}}(\boldsymbol{\theta})]$. Let π satisfy the same moment constraints as $\pi_{\boldsymbol{\lambda}}$, namely $\mathbb{E}_{\pi}[\mathbf{T}(\boldsymbol{\theta})] = \boldsymbol{\mu}(\boldsymbol{\lambda})$. Then

$$\mathbb{E}_{\pi}[\eta_{\boldsymbol{\lambda}}(\boldsymbol{\theta})] = \langle \boldsymbol{\lambda}, \mathbb{E}_{\pi}[\mathbf{T}(\boldsymbol{\theta})] \rangle = \langle \boldsymbol{\lambda}, \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\mathbf{T}(\boldsymbol{\theta})] \rangle = \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\eta_{\boldsymbol{\lambda}}(\boldsymbol{\theta})]. \tag{12}$$

Hence $\Omega(\pi) \geq \Omega(\pi_{\boldsymbol{\lambda}})$, which proves optimality. If Ω is strictly convex, this minimizer is unique. □

In Prop. 2 the infinite dimensional optimizer is certified by a finite dimensional potential lying in the span of the constraint functions. This motivates the induced family

$$\mathcal{Q}_{\Omega,\mathbf{T}} = \left\{ \pi_{\boldsymbol{\lambda}} \in \mathcal{P} : \pi_{\boldsymbol{\lambda}} \text{ attains } \Omega^*(\eta_{\boldsymbol{\lambda}}), \boldsymbol{\lambda} \in \mathbb{R}^P \right\}.$$

Props. 1 and 2 are the only occasions where the full density-space duality is needed. Once the induced family $Q_{\Omega,\mathbf{T}}$ and the potential Ψ^* are identified, all subsequent geometry and updates are finite-dimensional in the dual coordinates $(\boldsymbol{\lambda}, \boldsymbol{\mu})$. From here on, π_λ denotes a member of $\mathcal{Q}_{\Omega,\mathbf{T}}$. Whenever the maximizer is unique and the conjugate is differentiable at $\eta_{\boldsymbol{\lambda}}$, we write $\pi_{\boldsymbol{\lambda}} = \nabla\Omega^*(\eta_{\boldsymbol{\lambda}})$. When Ψ^* is differentiable at $\boldsymbol{\lambda}$ and the corresponding optimizer $\pi_{\boldsymbol{\lambda}}$ has finite moments,

$$\boldsymbol{\mu}(\boldsymbol{\lambda}) = \nabla_{\boldsymbol{\lambda}}\Psi^*(\boldsymbol{\lambda}) = \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\mathbf{T}(\boldsymbol{\theta})].$$

We further write $\boldsymbol{\lambda}(\boldsymbol{\mu}) = \nabla_{\boldsymbol{\mu}}\Psi(\boldsymbol{\mu})$. Thus $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ play the roles of canonical and expectation parameters, as in exponential families, but for the larger class of maxent families induced by Ω.

Finally, whenever $\pi_{\boldsymbol{\lambda}} \in \mathcal{Q}_{\Omega,\mathbf{T}}$ solves Eq. 11 at $\boldsymbol{\mu}$, the FYD admits the mixed-space form

$$\mathbb{D}_{\Omega}(\eta_{\boldsymbol{\lambda}_0}, \pi_{\boldsymbol{\lambda}}) = \Psi^*(\boldsymbol{\lambda}_0) + \Psi(\boldsymbol{\mu}) - \langle \boldsymbol{\lambda}_0, \boldsymbol{\mu} \rangle.$$

Indeed, Prop. 1 gives $\Omega^*(\eta_{\boldsymbol{\lambda}_0}) = \Psi^*(\boldsymbol{\lambda}_0)$, optimality of $\pi_{\boldsymbol{\lambda}}$ in Eq. 11 gives $\Omega(\pi_{\boldsymbol{\lambda}}) = \Psi(\boldsymbol{\mu})$, and $\mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\eta_{\boldsymbol{\lambda}_0}(\boldsymbol{\theta})] = \langle \boldsymbol{\lambda}_0, \boldsymbol{\mu} \rangle$. This is directly analogous to the mixed-space Bregman form of the KLD in Eq. 1 and enables BLR-style additive updates in canonical coordinates. Importantly, the first argument is now a reference potential $\eta_{\boldsymbol{\lambda}_0}$, not necessarily a normalized density. When $\Omega = -\mathcal{H}$ we recover the classical exponential family. When $\Omega = \Omega_{2-q}$, the induced family is the deformed q-exponential (proven in App. B.2). In App. F.1, we show the quadratic $\mathbf{T}(\boldsymbol{\theta})^\top = (\boldsymbol{\theta}, \boldsymbol{\theta}\boldsymbol{\theta}^\top)$ yields the bounded-support and heavy-tailed q-Gaussian families used by the q-VON Algorithms 1 and 2.

3.3 Generalizing the Variational Objective

Using FYDs and the generalized maxent families, the generalized BLR objective

$$\min_{\boldsymbol{\lambda} \in \Lambda} \mathbb{E}_{\boldsymbol{\theta} \sim \pi_{\boldsymbol{\lambda}}}[\ell(\boldsymbol{\theta})] + \mathbb{D}_{\Omega}(\eta_{\boldsymbol{\lambda}_0}, \pi_{\boldsymbol{\lambda}}), \quad \Lambda = \left\{ \boldsymbol{\lambda} \in \mathbb{R}^P : \arg\max_{\pi \in \mathcal{P}} \left(\mathbb{E}_{\pi}[\eta_{\boldsymbol{\lambda}}(\boldsymbol{\theta})] - \Omega(\pi) \right) \neq \varnothing \right\}, \tag{13}$$

recovers the BLR objective when $\Omega = -\mathcal{H}$ and $\eta_{\boldsymbol{\lambda}_0} = \log \pi_0$. Other choices of Ω amount to new objectives. From the definition of FYDs we can write Eq. 13 as

$$\min_{\boldsymbol{\lambda} \in \Lambda} \mathbb{E}_{\boldsymbol{\theta} \sim \pi_{\boldsymbol{\lambda}}}[\ell(\boldsymbol{\theta}) - \eta_{\boldsymbol{\lambda}_0}(\boldsymbol{\theta})] + \Omega(\pi_{\boldsymbol{\lambda}}). \tag{14}$$

This separation in terms of a linear expectation of $\eta_{\boldsymbol{\lambda}_0}$ and a generalized negentropy of $\pi_{\boldsymbol{\lambda}}$ is not possible with f-divergences (except the KLD). Up to a $\pi_{\boldsymbol{\lambda}}$-independent scalar Eq. 13 is the FYD $\mathbb{D}_{\Omega}(\eta_{\boldsymbol{\lambda}_0} - \ell, \pi_{\boldsymbol{\lambda}})$. Thus, observing a loss ℓ tilts the reference potential from $\eta_{\boldsymbol{\lambda}_0}$ to $\eta_{\boldsymbol{\lambda}_0} - \ell$. Since losses compose associatively in dual space, processing losses sequentially or jointly results in the same potential, preserving sequential coherence in score-space.

3.4 Generalizing the BLR

Theorem 1 (Fenchel-Young BLR). *Let $\mathcal{Q}_{\Omega,\mathbf{T}}$, Ψ, Ψ^*, and the moment map $\boldsymbol{\mu}(\boldsymbol{\lambda}) = \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\mathbf{T}(\boldsymbol{\theta})]$ be as in Sec. 3.2 and Props. 1 and 2. Let Λ_0 be an open coordinate chart of the induced family $\mathcal{Q}_{\Omega,\mathbf{T}}$. Assume that $\Psi^* \in C^2(\Lambda_0)$, and define a generalized Fisher[8] $\mathbf{F}(\boldsymbol{\lambda}) := \nabla^2_{\boldsymbol{\lambda}} \Psi^*(\boldsymbol{\lambda})$ invertible on Λ_0. Assume also that the dual potential Ψ is differentiable on $\boldsymbol{\mu}(\Lambda_0)$, with $\nabla_{\boldsymbol{\mu}} \Psi(\boldsymbol{\mu}(\boldsymbol{\lambda})) = \boldsymbol{\lambda}$. Fix $\boldsymbol{\lambda}_0 \in \Lambda_0$, and define $\bar{\ell}(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - \eta_{\boldsymbol{\lambda}_0}(\boldsymbol{\theta})$. Assume that the map $\boldsymbol{\mu} \mapsto \mathbb{E}_{\pi_{\boldsymbol{\lambda}(\boldsymbol{\mu})}}[\bar{\ell}(\boldsymbol{\theta})]$ is differentiable on $\boldsymbol{\mu}(\Lambda_0)$. Then the generalized canonical gradient satisfies*

$$\mathbf{F}(\boldsymbol{\lambda})^{-1} \nabla_{\boldsymbol{\lambda}} \left[\mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\ell(\boldsymbol{\theta})] + \mathbb{D}_{\Omega}(\eta_{\boldsymbol{\lambda}_0}, \pi_{\boldsymbol{\lambda}}) \right] = \nabla_{\boldsymbol{\mu}} \mathbb{E}_{\pi_{\boldsymbol{\lambda}(\boldsymbol{\mu})}}[\bar{\ell}(\boldsymbol{\theta})] \Big|_{\boldsymbol{\mu}=\boldsymbol{\mu}(\boldsymbol{\lambda})} + \boldsymbol{\lambda}. \tag{15}$$

Consequently, the canonical-gradient step is

$$\begin{aligned} \boldsymbol{\lambda}_{t+1} &\leftarrow \boldsymbol{\lambda}_t - \rho_t \mathbf{F}(\boldsymbol{\lambda}_t)^{-1} \nabla_{\boldsymbol{\lambda}} \left(\mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\ell(\boldsymbol{\theta})] + \mathbb{D}_{\Omega}(\eta_{\boldsymbol{\lambda}_0}, \pi_{\boldsymbol{\lambda}}) \right) \Big|_{\boldsymbol{\lambda}_t} \\ &= (1-\rho_t)\boldsymbol{\lambda}_t - \rho_t \nabla_{\boldsymbol{\mu}} \mathbb{E}_{\pi_{\boldsymbol{\lambda}(\boldsymbol{\mu})}}[\bar{\ell}(\boldsymbol{\theta})] \Big|_{\boldsymbol{\mu}_t}. \end{aligned} \tag{16}$$

In particular, any fixed point with $\rho_t > 0$ satisfies $\boldsymbol{\lambda}^* = -\nabla_{\boldsymbol{\mu}} \mathbb{E}_{\pi_{\boldsymbol{\lambda}(\boldsymbol{\mu})}}[\bar{\ell}(\boldsymbol{\theta})] \Big|_{\boldsymbol{\mu}=\boldsymbol{\mu}(\boldsymbol{\lambda}^*)}$.

The proof is in App. C. The canonical parameter at convergence is matched with the gradient of the loss. Analogous to the fixed-point interpretation of the BLR, this shows that generalized canonical gradients are intrinsic to the variational problem, rather than an optimization choice. The update in Eq. 16 is one of our main contributions. We call it the **generalized BLR**, as it applies to families of distributions larger than the exponential family (see Fig. 1). When $\Omega = -\mathcal{H}$ we recover the BLR. If $\eta_{\boldsymbol{\lambda}_0}(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}) = \eta_{\boldsymbol{\lambda}'}(\boldsymbol{\theta}) + \text{const.}$, then $\boldsymbol{\lambda}_{t+1} = (1-\rho_t)\boldsymbol{\lambda}_t + \rho_t \boldsymbol{\lambda}'$, so a unit step recovers the generalized conjugate posterior induced by the tilted potential.

4 Example: Tsallis Negentropy and q-VON

As our main example we replace the Shannon negentropy by the Tsallis negentropy Ω_{2-q} and derive Algorithms 1 and 2 (see App. F for details), which we apply to CIFAR-100, CIFAR-10, MNIST (see App. G), and to Bayesian logistic regression on a UCI dataset (Fig. 2). The induced maxent family is the q-exponential family (see App. B.2 for a proof),

$$\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \exp_q \left(\boldsymbol{\lambda}^\top \mathbf{T}(\boldsymbol{\theta}) - A_{2-q}(\boldsymbol{\lambda}) \right), \tag{17}$$

where $\exp_q(u) = [1 + (1-q)u]_+^{1/(1-q)}$ when $q \neq 1$. We take $\Theta = \mathbb{R}^D$, with Lebesgue reference measure, so $\boldsymbol{\theta} \in \mathbb{R}^D$. For the quadratic statistic $\mathbf{T}(\boldsymbol{\theta})^\top = (\boldsymbol{\theta}, \boldsymbol{\theta}\boldsymbol{\theta}^\top)$ with canonical parameter $\boldsymbol{\lambda}^\top = (\mathbf{S}\mathbf{m}, -\frac{1}{2}\mathbf{S})$, this yields a location-scale $(\mathbf{m}, \boldsymbol{\Sigma})$ family of q-Gaussians. The quadratic geometry is unchanged, but q controls support and tail shapes: $q < 1$ gives parameter-dependent bounded-support Alg. 1, and $q > 1$ gives heavy-tailed Alg. 2 with $\nu = 2/(q-1) - D$. In App. D we review the canonical, location-scale, and expectation parametrizations of bounded-support and heavy-tailed q-Gaussians, as well as a closed-form FYD for the heavy-tailed case, which we use to evaluate the objective optimized by Alg. 2. In App. E we derive the reparameterized gradient estimators underlying lines 2-3 of Algorithms 1 and 2.

Specializing the generalized BLR update (16) to quadratic families with expectation parameters $(\boldsymbol{\mu}, \mathbf{M}) = (\mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\boldsymbol{\theta}], \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\boldsymbol{\theta}\boldsymbol{\theta}^\top])$, we obtain Newton-like canonical updates. After converting gradients with respect to $(\boldsymbol{\mu}, \mathbf{M})$ to gradients with respect to $(\mathbf{m}, \boldsymbol{\Sigma})$ and applying Bonnet- and Price-type identities (App. E), both the bounded-support and heavy-tailed cases take the common form (App. F)

$$\mathbf{S}_{t+1} = (1-\rho_t)\mathbf{S}_t + \rho_t \mathbb{E}_{\tilde{\pi}_t^{2-q}} \left[\nabla^2_{\boldsymbol{\theta}} \bar{\ell}(\boldsymbol{\theta}) \right], \quad \mathbf{m}_{t+1} = \mathbf{m}_t - \rho_t \mathbf{S}_{t+1}^{-1} \mathbb{E}_{\pi_t} \left[\nabla_{\boldsymbol{\theta}} \bar{\ell}(\boldsymbol{\theta}) \right], \tag{18}$$

where $\tilde{\pi}^{2-q}$ denotes the $(2-q)$-escort of π. The updates in Eq. 18 are strikingly similar in form to those of VON and IVON [Khan et al., 2018, Osawa et al., 2019, Shen et al., 2024]. The difference

[8] Note that this is different from the "q-Fisher" discussed by Amari et al. [2012] who consider instead the Hessian of the cumulant of the solution to an escort-expectation constrained variant of Eq. 11, which leads to a different geometry than ours when $q \neq 1$.

Algorithm 1 q-VON (bounded-support $q < 1$)

Require: Learning rates ρ_t, weight-decay $\delta > 0$
Require: Momentum $\beta_1, \beta_2 \in [0, 1)$
Require: Hessian init $h_0 > 0$,
Deformation index $q < 1$
Compute: $\kappa(q, D) \leftarrow \frac{R^2}{D+2(2-q)/(1-q)}$,
$\gamma \leftarrow D + 2/(1-q)$
Init: $\mathbf{m} \leftarrow$ (NN-weights), $\mathbf{h} \leftarrow h_0$, $\mathbf{g} \leftarrow 0$, $\zeta \leftarrow N$
Init: $\mathbf{s} \leftarrow \zeta(\mathbf{h} + \delta)$ (diagonal of $\mathbf{S}$)
Init: $\boldsymbol{\sigma}^2 \leftarrow (s_1 \cdot \ldots \cdot s_D)^{1/\gamma} / \mathbf{s}$ (scale)

for $t = 1, 2, \ldots$ **do**
  $\widehat{\mathbf{g}} \leftarrow \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$, $\quad \boldsymbol{\theta} \sim \mathcal{N}_q(\mathbf{m}, \mathrm{diag}(\boldsymbol{\sigma}^2))$
  $\widehat{\mathbf{h}} \leftarrow \kappa^{-1} \widehat{\mathbf{g}} \cdot (\boldsymbol{\theta} - \mathbf{m}) / \boldsymbol{\sigma}^2$
  (estimator for $\mathbb{E}_{\tilde{\pi}}[\mathrm{diag}(\nabla^2 \ell)]$)
  $\mathbf{g} \leftarrow \beta_1 \mathbf{g} + (1 - \beta_1) \widehat{\mathbf{g}}$
  $\mathbf{h} \leftarrow \beta_2 \mathbf{h} + (1 - \beta_2) \widehat{\mathbf{h}}$
  $\mathbf{m} \leftarrow \mathbf{m} - \rho_t (\mathbf{g} + \delta \mathbf{m}) / (\mathbf{h} + \delta)$
  $\mathbf{s} \leftarrow \zeta(\mathbf{h} + \delta)$
  $\boldsymbol{\sigma}^2 \leftarrow (s_1 \cdot \ldots \cdot s_D)^{1/\gamma} / \mathbf{s}$
end for
return $\mathbf{m}, \boldsymbol{\sigma}^2$

Algorithm 2 Student-t-VON (heavytailed $q > 1$)

Require: Learning rates ρ_t, weight-decay $\delta > 0$
Require: Momentum $\beta_1, \beta_2 \in [0, 1)$
Require: Hessian init $h_0 > 0$,
Degrees of freedom $\nu > 2$
Compute:
$$\kappa_t \leftarrow \frac{\nu + D}{\nu} \left[(\nu + D)^{\frac{D}{2}} \pi^{\frac{D}{2}} \frac{\Gamma(\frac{\nu}{2})}{\Gamma(\frac{\nu+D}{2})} \right]^{\frac{2}{\nu}}$$
Init: $\mathbf{m} \leftarrow$ (NN-weights), $\mathbf{h} \leftarrow h_0$, $\mathbf{g} \leftarrow 0$, $\zeta \leftarrow N$
Init: $\mathbf{s} \leftarrow \zeta(\mathbf{h} + \delta)$ (diagonal of $\mathbf{S}$)
Init: $\boldsymbol{\sigma}^2 \leftarrow \kappa_t (s_1 \cdot \ldots \cdot s_D)^{-1/\nu} / \mathbf{s}$ (scale)

for $t = 1, 2, \ldots$ **do**
  $\widehat{\mathbf{g}} \leftarrow \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$, $\quad \boldsymbol{\theta} \sim \mathcal{T}_\nu(\mathbf{m}, \mathrm{diag}(\boldsymbol{\sigma}^2))$
  $\widehat{\mathbf{h}} \leftarrow \frac{\nu-2}{\nu} \widehat{\mathbf{g}} \cdot (\boldsymbol{\theta} - \mathbf{m}) / \boldsymbol{\sigma}^2$
  (estimator for $\mathbb{E}_{\tilde{\pi}}[\mathrm{diag}(\nabla^2 \ell)]$)
  $\mathbf{g} \leftarrow \beta_1 \mathbf{g} + (1 - \beta_1) \widehat{\mathbf{g}}$
  $\mathbf{h} \leftarrow \beta_2 \mathbf{h} + (1 - \beta_2) \widehat{\mathbf{h}}$
  $\mathbf{m} \leftarrow \mathbf{m} - \rho_t (\mathbf{g} + \delta \mathbf{m}) / (\mathbf{h} + \delta)$
  $\mathbf{s} \leftarrow \zeta(\mathbf{h} + \delta)$
  $\boldsymbol{\sigma}^2 \leftarrow \kappa_t (s_1 \cdot \ldots \cdot s_D)^{-1/\nu} / \mathbf{s}$
end for
return $\mathbf{m}, \boldsymbol{\sigma}^2$

between $q < 1$ and $q > 1$ lies in the relation between $(\mathbf{S}, \mathbf{m})$ and the location-scale parametrizations, and in the corresponding escort reweighting appearing in the Price identity. The complete derivations, including the bounded-support and Student-t closed forms, and the escort-based Hessian identities are given in Apps. D to F. Algorithms 1 and 2 give practical diagonal implementations of these updates. Relative to IVON, only the sampling law, the constants relating $\mathbf{S}$ to the scale, and the corresponding diagonal Stein estimator change. As in IVON, we use minibatch stochastic gradients, a diagonal precision $\mathbf{S} = \mathrm{diag}(\mathbf{s})$, the reparametrization $\mathbf{s} = \zeta(\mathbf{h} + \delta)$, and momentum terms for stable deep-learning optimization. While the CIFAR-100 experiment we show in App. G is not a large-scale benchmark, there is no structural obstacle to scaling further, and applying q-VON to transformer- or LLM-scale training is an empirical engineering question, which we leave to future work.

Mean-Field. In high-dimensional models we use a factorized mean-field q-VON-MF, where each parameter follows a 1D q-Gaussian. We replace Ω in Eq. 13 with a sum of 1D terms. The only change to Algorithms 1 and 2 is to use 1D constants (set $D = 1$), and update each scale parameter separately rather than through the product $s_1 \cdots s_D$ (App. F.2). This addresses the problem that finite covariance in high-dimensional settings forces heavy-tailed q-values to be close to 1, and that D can dominate the max support radius in bounded-support settings. An exciting future direction that our work opens is training deep networks whose parameters have variable levels of stochasticity; for example, each layer may be associated with distinct, learnable or tunable support- and tail-behaviors.

5 Experiments

We evaluate q-VON on Bayesian logistic regression and image classification. Detailed experiment descriptions are in App. G, and results on varying q are in App. H. On Bayesian logistic regression, shrinking the max support radius by decreasing $q < 1$ lets the q-VON approximation closely track the MCMC posterior mode, in contrast to the Gaussian approximation (Fig. 2). On MNIST, CIFAR-10, and CIFAR-100 we compare q-VON and q-VON-MF against SGD, VSGD, IVON, and a mean-only baseline q-VSGD, using test accuracy, NLL, ECE, Brier score, and AUROC. Generally, q-VON improves over SGD and q-VSGD, showing that learning the posterior shape outperforms noise injection. The choice of tail behavior can trade off generalization and calibration: on MNIST, heavy-tailed q-VON has the highest accuracy while bounded-support q-VON has the best calibration.

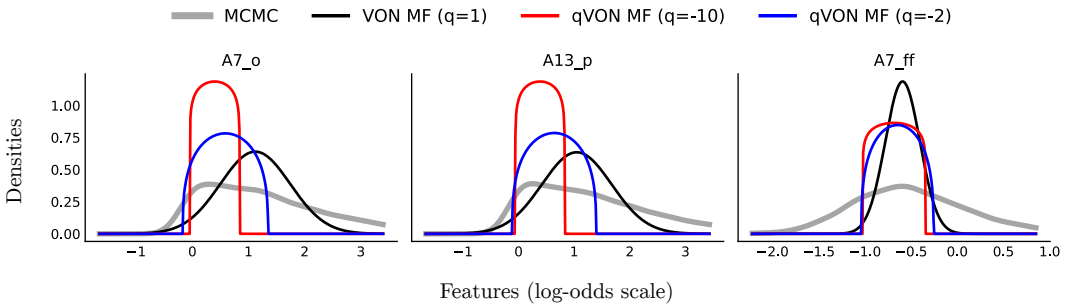


Figure 2: Bounded-support, mean-field variational posterior (red, blue) means can approach the marginal posterior modes of the coefficients associated with features A7_0, A13_p and A7_ff, by changing the value of q in Bayesian logistic regression on the UCI Credit Approval dataset. This is useful when posterior marginals (grey) are skewed, and thus the empirical mode is badly approximated by a Gaussian posterior (black).

On CIFAR, q-VON-MF is competitive with (highly tuned) IVON without additional tuning. Our experiments demonstrate that generalized BLR updates can be implemented at the scale of modern neural-network optimizers while extending IVON to posterior families unavailable to the BLR.

6 Discussion

We show how the simple additive updates of Khan and Rue [2023]'s BLR can be extended beyond classical Bayesian updating of exponential family posteriors. This reveals the deeper nature of these updates, and Bayes' rule itself, whose foundations are built on the duality between spaces of functions and measures. A general convex function defines a map between these spaces, with the Shannon entropy recovering the classical Bayesian structure. In practice, this insight allows us to generalize state-of-the-art variational online Newton (VON) methods to non-Gaussian posteriors by changing only a few lines of code. Our experiments suggest potential advantages, such as approaching the MAP estimate while still capturing uncertainty.

As the BLR covers a large variety of methods other than VON (see Khan and Rue [2023]), our generalization provides a basis for extending these methods beyond KL geometry. For instance, EVON [Minut et al., 2026] runs IVON in the eigenbasis of SOAP's preconditioner. An interesting direction for future work is to derive corresponding updates for mean-field q-Gaussian posteriors in this basis. In particular, a mean-field Student-t posterior is not invariant to rotation even when its marginal variances are equal, so the learned rotation can change the joint posterior while leaving its covariance unchanged. Beyond the Tsallis negative entropies we analyze here, convex generators constructed from proper scoring rules [Dawid, 2007] on the parameter space allow us to derive useful posterior families as unrestricted optimizers of appropriate variational objectives. For example, radial posteriors derived in this way can lead to a geometry-aware alternative to Farquhar et al. [2020].

Acknowledgements

We thank Kenichi Bannai for referring us to Lasota and Mackey [2013]'s proof.

This work is supported by the Bayes duality project, JST CREST Grant Number JPMJCR2112. Thomas Moellenhoff acknowledges the support of JSPS KAKENHI Grant Number 26H02541. This work is partially funded by FCT - Fundação para a Ciência e a Tecnologia, I.P., and, when eligible, co-funded by EU funds under project/support UID/50008/2025 - Instituto de Telecomunicações, with DOI identifier https://doi.org/10.54499/UID/50008/2025. This work is partially funded by the project DECOLLAGE (ERC-2022-CoG 101088763).

References

P. Alquier. Non-Exponentially Weighted Aggregation: Regret Bounds for Unbounded Loss Functions. In *International Conference on Machine Learning (ICML)*, 2021.

S.-I. Amari. α-Divergence Is Unique, Belonging to both f-divergence and Bregman Divergence Classes. *IEEE Transactions on Information Theory*, 55(11):4925–4931, 2009.

S.-I. Amari. *Information Geometry and its Applications*, volume 194. Springer, 2016.

S.-I. Amari and A. Cichocki. Information geometry of divergence functions. *Bulletin of the Polish academy of sciences. Technical Sciences*, 58(1):183–195, 2010.

S.-I. Amari and A. Ohara. Geometry of q-Exponential Family of Probability Distributions. *Entropy*, 13(6):1170–1185, 2011.

S.-I. Amari, A. Ohara, and H. Matsuzoe. Geometry of deformed exponential families: Invariant, dually-flat and conformal geometries. *Physica A: Statistical Mechanics and its Applications*, 391 (18):4308–4319, 2012.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A General Framework for Updating Belief Distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

M. Blondel, A. Martins, and V. Niculae. Learning classifiers with Fenchel-Young losses: Generalized entropies, margins, and algorithms. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 606–615. PMLR, 2019.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

O. Catoni. PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning. *arXiv preprint arXiv:0712.0248*, 2007.

A. P. Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59(1):77–93, 2007.

S. Farquhar, M. A. Osborne, and Y. Gal. Radial Bayesian neural networks: Beyond discrete support in large-scale Bayesian deep learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

A. Ghosh and A. Basu. Robust Bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, 68(2):413–437, 2016.

J. W. Gibbs. *Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundations of Thermodynamics*. C. Scribner's sons, 1902.

P. D. Grünwald and N. A. Mehta. Fast Rates for General Unbounded Loss Functions: From ERM to Generalized Bayes. *Journal of Machine Learning Research*, 21(56):1–80, 2020.

T. Guilmeau, E. Chouzenoux, and V. Elvira. Regularized rényi divergence minimization through bregman proximal gradient algorithms. *Journal of Machine Learning Research*, 26(157):1–56, 2025. URL https://www.jmlr.org/papers/v26/23-0573.html.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

J. Hernandez-Lobato, Y. Li, M. Rowland, T. Bui, D. Hernández-Lobato, and R. Turner. Black-box α-divergence Minimization. In *International conference on machine learning*, pages 1511–1520. PMLR, 2016.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic Variational Inference. *Journal of machine learning research*, 2013.

P. Ishwar and P. Moulin. On the existence and characterization of the maxent distribution under general moment inequality constraints. *IEEE transactions on information theory*, 51(9):3322–3333, 2005.

E. T. Jaynes. Information Theory and Statistical Mechanics. *Physical review*, 106(4):620, 1957.

J. Jewson, J. Q. Smith, and C. Holmes. Principles of Bayesian Inference Using General Divergence Criteria. *Entropy*, 20(6):442, 2018.

M. Khan and W. Lin. Conjugate-Computation Variational Inference: Converting Variational Inference in Non-conjugate Models to Inferences in Conjugate Models. In *Artificial Intelligence and Statistics*, pages 878–887. PMLR, 2017.

M. E. Khan and H. Rue. The Bayesian Learning Rule. *Journal of Machine Learning Research*, 24 (281):1–46, 2023.

M. E. Khan, D. Nielsen, V. Tangkaratt, W. Lin, Y. Gal, and A. Srivastava. Fast and Scalable Bayesian Deep Learning by Weight-Perturbation in Adam. In *International Conference on Machine Learning (ICML)*, 2018.

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):1–109, 2022.

Z. Landsman and J. Nešlehová. Stein's lemma for elliptical random vectors. *Journal of Multivariate Analysis*, 99(5):912–927, 2008.

Z. Landsman, S. Vanduffel, and J. Yao. Some Stein-type Inequalities for Multivariate Elliptical Distributions and Applications. *Statistics & Probability Letters*, 97:54–62, 2015.

A. Lasota and M. C. Mackey. *Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics*, volume 97. Springer Science & Business Media, 2013.

Y. Li and R. E. Turner. Rényi Divergence Variational Inference. *Advances in neural information processing systems*, 29, 2016.

M. Lukashchuk, R. Trésor, W. Nuijten, I. Senoz, and B. Vries. The Quotient Bayesian Learning Rule. *Advances in Neural Information Processing Systems*, 38:8788–8827, 2026.

A. F. Martins, M. Treviso, A. Farinhas, P. M. Aguiar, M. A. Figueiredo, M. Blondel, and V. Niculae. Sparse continuous distributions and Fenchel-Young losses. *Journal of Machine Learning Research*, 23(257):1–74, 2022.

H. Matsuzoe and A. Ohara. Geometry for q-exponential families. *Recent progress in differential geometry and its related fields*, pages 55–71, 2011.

J. W. Miller and D. B. Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 2019.

T. Minka et al. Divergence Measures and Message Passing. Technical report, Microsoft Research, 2005.

A. R. Minut, N. Daheim, M. Miani, M. E. Khan, W. Lin, and T. Möllenhoff. Soap-bubbles: Structured weight uncertainty for neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026.

K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT press, 2012.

J. Naudts. Deformed exponentials and logarithms in generalized thermostatistics. *Physica A: Statistical Mechanics and its Applications*, 316(1-4):323–334, 2002.

J. Naudts. *Generalised Thermostatistics*. Springer Science & Business Media, 2011.

F. Nielsen. Statistical Divergences between Densities of Truncated Exponential Families with Nested Supports: Duo Bregman and duo Jensen divergences. *Entropy*, 24(3):421, 2022.

M. Opper and C. Archambeau. The Variational Gaussian Approximation Revisited. *Neural computation*, 21(3):786–792, 2009.

A. Orvieto, H. Kersting, F. Proske, F. Bach, and A. Lucchi. Anticorrelated Noise Injection for Improved Generalization. In *International Conference on Machine Learning (ICML)*, 2022.

K. Osawa, S. Swaroop, M. E. Khan, A. Jain, R. Eschenhagen, R. E. Turner, and R. Yokota. Practical deep learning with Bayesian principles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Y. Shen, N. Daheim, B. Cong, P. Nickl, G. M. Marconi, C. Bazan, R. Yokota, I. Gurevych, D. Cremers, M. E. Khan, and T. Möllenhoff. Variational Learning is Effective for Large Deep Networks. In *International Conference on Machine Learning (ICML)*, 2024.

S. Sklaviadis, T. Möllenhoff, E. Khan, A. Martins, and M. Figueiredo. The Fenchel-Young Generalization of Bayesian Principles, 2025. URL https://arxiv.org/abs/2502.10295v4. arXiv:2502.10295v4 [cs.LG], revised 22 October 2026.

S. Sklaviadis, T. Möllenhoff, M. A. T. Figueiredo, A. F. T. Martins, and M. E. Khan. A Stein Identity for q-Gaussians with Bounded Support. In *Conference on Parsimony and Learning (CPAL)*, 2026.

H. Touchette. Legendre-Fenchel transforms in a nutshell. *URL http://www. maths. qmul. ac. uk/ht/archive/lfth2. pdf*, page 25, 2005.

M. J. Wainwright, M. I. Jordan, et al. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

P. M. Williams. Bayesian Conditionalisation and the Principle of Minimum Information. *The British Journal for the Philosophy of Science*, 31(2):131–144, 1980.

K. Wu and J. R. Gardner. Understanding Stochastic Natural Gradient Variational Inference. *arXiv preprint arXiv:2406.01870*, 2024.

A. Zellner. Optimal Information Processing and Bayes's Theorem. *The American Statistician*, 42(4): 278–280, 1988.

T. Zhang. Theoretical Analysis of a Class of Randomized Regularization Methods. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 156–163, 1999.

J. Zhu, N. Chen, and E. P. Xing. Bayesian Inference with Posterior Regularization and Applications to Infinite Latent SVMs. *The Journal of Machine Learning Research*, 15(1):1799–1847, 2014.

Appendix Contents

A The sharp Gibbs bound forces the KLD **15**

B Maxent **15**

B.1 Exponential family as max Shannon entropy 15

B.2 q-exponential family as Max Tsallis $(2-q)$-entropy 17

B.3 Existence 18

C Proof of Thm. 1 **19**

D q-Gaussians: q-exponential Family Form to Location-Scale **19**

D.1 Canonical q-exponential form 19

D.2 Radial laws and location-scale form 20

D.3 Expectation parameters 21

D.4 FY Mixed-space Bregman between two heavy-tailed q-Gaussians 22

E Bonnet-, Price-, and Stein-type identities for q-Gaussians **22**

E.1 Bonnet-type identities 23

E.2 Price-type identities 24

F Tsallis Negentropy and q-VON **26**

F.1 q-Variational Online Newton 27

F.2 Mean-Field Variants 30

G Experiments **30**

G.1 Bayesian logistic regression 30

G.2 Image Classification with Neural Networks 30

G.3 Empirical Verification of Convergence 32

H Varying q **33**

H.1 1D Logistic Regression 33

H.2 MNIST 34

A The sharp Gibbs bound forces the KLD

Proposition 3. *Fix a prior density π_0 with respect to h. Suppose $\mathbb{D}(\pi, \pi_0)$ is a loss-independent functional such that for every measurable loss $\ell : \Theta \to \mathbb{R} \cup \{+\infty\}$ with*

$$0 < \int e^{-\ell(\boldsymbol{\theta})} \pi_0(\boldsymbol{\theta}) dh(\boldsymbol{\theta}) < \infty$$

and every $\pi \in \mathcal{P}$,

$$-\log \int e^{-\ell(\boldsymbol{\theta})} \pi_0(\boldsymbol{\theta}) dh(\boldsymbol{\theta}) \leq \mathbb{E}_\pi[\ell(\boldsymbol{\theta})] + \mathbb{D}(\pi, \pi_0),$$

with equality at the Gibbs density

$$\pi_\ell(\boldsymbol{\theta}) = \frac{e^{-\ell(\boldsymbol{\theta})} \pi_0(\boldsymbol{\theta})}{\int e^{-\ell(\boldsymbol{\theta}')} \pi_0(\boldsymbol{\theta}') dh(\boldsymbol{\theta}')}.$$

Then for every $\pi \ll \pi_0$ such that $\mathbb{D}_{KL}(\pi \,\|\, \pi_0) < \infty$, defining

$$\ell_\pi(\boldsymbol{\theta}) = -\log \frac{\pi(\boldsymbol{\theta})}{\pi_0(\boldsymbol{\theta})},$$

we have

$$\mathbb{D}(\pi, \pi_0) = \mathbb{D}_{KL}(\pi \,\|\, \pi_0).$$

Proof. We choose a loss whose Gibbs posterior is the candidate posterior we want, and then evaluate the variational identity there. Let $\ell = \ell_\pi$. Then

$$\int e^{-\ell_\pi(\boldsymbol{\theta})} \pi_0(\boldsymbol{\theta}) dh(\boldsymbol{\theta}) = \int \pi(\boldsymbol{\theta}) dh(\boldsymbol{\theta}) = 1,$$

so the left-hand side of the inequality is 0. Hence

$$0 = \mathbb{E}_\pi \left[-\log \frac{\pi(\boldsymbol{\theta})}{\pi_0(\boldsymbol{\theta})} \right] + \mathbb{D}(\pi, \pi_0) = -\mathbb{D}_{\mathrm{KL}}(\pi \,\|\, \pi_0) + \mathbb{D}(\pi, \pi_0).$$

□

B Maxent

B.1 Exponential family as max Shannon entropy

The following proof is based on Lasota and Mackey [2013, Prop. 9.1.2].

Proposition 4. *Let $(\Theta, \mathcal{A}, h)$ be a measure space, and let*

$$\mathbf{T}(\boldsymbol{\theta}) = (T_1(\boldsymbol{\theta}), ..., T_P(\boldsymbol{\theta}))$$

be a measurable vector-valued function. Assume there exist

$$\boldsymbol{\mu} = (\mu_1, ..., \mu_P) \qquad \textit{and} \qquad \boldsymbol{\lambda} = (\lambda_1, ..., \lambda_P)$$

such that

$$\mu_i = \frac{\int_\Theta T_i(\boldsymbol{\theta}) \exp\left(\langle \boldsymbol{\lambda}, \mathbf{T}(\boldsymbol{\theta}) \rangle\right) dh(\boldsymbol{\theta})}{\int_\Theta \exp\left(\langle \boldsymbol{\lambda}, \mathbf{T}(\boldsymbol{\theta}) \rangle\right) dh(\boldsymbol{\theta})}, \qquad i = 1, ..., P,$$

and all of the integrals are finite. Then the maximum of the Shannon entropy

$$\mathcal{H}(\pi) = -\int_\Theta \pi(\boldsymbol{\theta}) \log \pi(\boldsymbol{\theta}) dh(\boldsymbol{\theta})$$

over all densities $\pi \in \mathcal{P}$ satisfying

$$\mu_i = \int_\Theta T_i(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) dh(\boldsymbol{\theta}), \qquad i = 1, ..., P,$$

is attained at

$$\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \frac{\exp\left(\langle \boldsymbol{\lambda}, \mathbf{T}(\boldsymbol{\theta}) \rangle\right)}{\int_\Theta \exp\left(\langle \boldsymbol{\lambda}, \mathbf{T}(\boldsymbol{\theta}) \rangle\right) dh(\boldsymbol{\theta})}.$$

Proof. Set

$$Z = \int_\Theta \exp\left(\langle \boldsymbol{\lambda}, \mathbf{T}(\boldsymbol{\theta})\rangle\right) dh(\boldsymbol{\theta}), \qquad \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = Z^{-1} \exp\left(\langle \boldsymbol{\lambda}, \mathbf{T}(\boldsymbol{\theta})\rangle\right).$$

Let $\pi \in \mathcal{P}$ satisfy

$$\mu_i = \int_\Theta T_i(\boldsymbol{\theta})\pi(\boldsymbol{\theta}) dh(\boldsymbol{\theta}), \qquad i = 1, ..., P.$$

Then

$$\begin{aligned}
\mathcal{H}(\pi) &= -\int_\Theta \pi(\boldsymbol{\theta}) \log \pi(\boldsymbol{\theta}) dh(\boldsymbol{\theta}) \\
&= -\int_\Theta \pi(\boldsymbol{\theta}) \log \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) dh(\boldsymbol{\theta}) - \int_\Theta \pi(\boldsymbol{\theta}) \log \frac{\pi(\boldsymbol{\theta})}{\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})} dh(\boldsymbol{\theta}) \\
&\leq -\int_\Theta \pi(\boldsymbol{\theta}) \log \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) dh(\boldsymbol{\theta}),
\end{aligned}$$

since

$$\int_\Theta \pi(\boldsymbol{\theta}) \log \frac{\pi(\boldsymbol{\theta})}{\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})} dh(\boldsymbol{\theta}) \geq 0.$$

Now

$$\begin{aligned}
-\int_\Theta \pi(\boldsymbol{\theta}) \log \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\, dh(\boldsymbol{\theta}) &= -\int_\Theta \pi(\boldsymbol{\theta}) \left(\langle \boldsymbol{\lambda}, \mathbf{T}(\boldsymbol{\theta})\rangle - \log Z\right) dh(\boldsymbol{\theta}) \\
&= \log Z - \sum_{i=1}^P \lambda_i \int_\Theta T_i(\boldsymbol{\theta})\pi(\boldsymbol{\theta}) dh(\boldsymbol{\theta}) \\
&= \log Z - \sum_{i=1}^P \lambda_i \mu_i.
\end{aligned}$$

Therefore,

$$\mathcal{H}(\pi) \leq \log Z - \sum_{i=1}^P \lambda_i \mu_i.$$

It remains to show that $\pi_{\boldsymbol{\lambda}}$ attains this value. We compute

$$\begin{aligned}
\mathcal{H}(\pi_{\boldsymbol{\lambda}}) &= -\int_\Theta \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) \log \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) dh(\boldsymbol{\theta}) \\
&= -\int_\Theta \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) \left(\langle \boldsymbol{\lambda}, \mathbf{T}(\boldsymbol{\theta})\rangle - \log Z\right) dh(\boldsymbol{\theta}) \\
&= \log Z - \sum_{i=1}^P \lambda_i \int_\Theta T_i(\boldsymbol{\theta})\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) dh(\boldsymbol{\theta}).
\end{aligned}$$

By the definition of $\pi_{\boldsymbol{\lambda}}$,

$$\int_\Theta T_i(\boldsymbol{\theta})\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) dh(\boldsymbol{\theta}) = \frac{\int_\Theta T_i(\boldsymbol{\theta}) \exp\left(\langle \boldsymbol{\lambda}, \mathbf{T}(\boldsymbol{\theta})\rangle\right) dh(\boldsymbol{\theta})}{\int_\Theta \exp\left(\langle \boldsymbol{\lambda}, \mathbf{T}(\boldsymbol{\theta})\rangle\right) dh(\boldsymbol{\theta})} = \mu_i.$$

Hence

$$\mathcal{H}(\pi_{\boldsymbol{\lambda}}) = \log Z - \sum_{i=1}^P \lambda_i \mu_i.$$

So for every feasible π,

$$\mathcal{H}(\pi) \leq \mathcal{H}(\pi_{\boldsymbol{\lambda}}),$$

which proves that $\pi_{\boldsymbol{\lambda}}$ maximizes the entropy. □

B.2 q-exponential family as Max Tsallis $(2-q)$-entropy

Proposition 5. *Let $(\Theta, \mathcal{A}, h)$ be a measure space, and let*

$$\mathbf{T}(\boldsymbol{\theta}) = (T_1(\boldsymbol{\theta}), ..., T_P(\boldsymbol{\theta}))$$

be measurable. Fix $q < 2$, $q \neq 1$, and define

$$\beta_q(t) = \begin{cases} \dfrac{t^{2-q} - t}{(2-q)(1-q)}, & t \geq 0, \\ +\infty, & t < 0 \end{cases}$$

and

$$\Omega_{2-q}(\pi) = \int_\Theta \beta_q(\pi(\boldsymbol{\theta}))dh(\boldsymbol{\theta}).$$

On the simplex of densities,

$$\Omega_{2-q}(\pi) = -\frac{1}{2-q}\mathcal{S}_{2-q}(\pi), \qquad \mathcal{S}_{2-q}(\pi) = \frac{1 - \int_\Theta \pi(\boldsymbol{\theta})^{2-q} dh(\boldsymbol{\theta})}{1-q}.$$

Assume there exist $\boldsymbol{\mu} = (\mu_1, ..., \mu_P)$ and $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_P)$ such that

$$\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \exp_q\left(\langle \boldsymbol{\lambda}, \mathbf{T}(\boldsymbol{\theta})\rangle - A_{2-q}(\boldsymbol{\lambda})\right)$$

is a density, $\Omega_{2-q}(\pi_{\boldsymbol{\lambda}}) < \infty$, all the integrals below are finite, and

$$\mu_i = \int_\Theta T_i(\boldsymbol{\theta})\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})dh(\boldsymbol{\theta}), \qquad i = 1, ..., P.$$

Then $\pi_{\boldsymbol{\lambda}}$ minimizes Ω_{2-q}, equivalently maximizes $\mathcal{S}_{2-q}$, over all densities $\pi \in \mathcal{P}$ satisfying

$$\mu_i = \int_\Theta T_i(\boldsymbol{\theta})\pi(\boldsymbol{\theta})dh(\boldsymbol{\theta}), \qquad i = 1, ..., P.$$

Proof. Set

$$\eta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \langle \boldsymbol{\lambda}, \mathbf{T}(\boldsymbol{\theta})\rangle - A_{2-q}(\boldsymbol{\lambda}).$$

Let $\pi \in \mathcal{P}$ satisfy the same moment constraints as $\pi_{\boldsymbol{\lambda}}$.

We claim that for h-a.e. $\boldsymbol{\theta}$,

$$\eta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) + \frac{1}{2-q} \in \partial\beta_q\left(\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\right).$$

If $\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) > 0$, then

$$\log_q \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \eta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}),$$

and since

$$\beta_q'(t) = \log_q(t) + \frac{1}{2-q}, \qquad t > 0,$$

it follows that

$$\eta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) + \frac{1}{2-q} = \beta_q'\left(\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\right) \in \partial\beta_q\left(\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\right).$$

If $q < 1$ and $\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = 0$, then by the definition of $\exp_q$ we have

$$\eta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) \leq -\frac{1}{1-q}.$$

Also, the right derivative at 0 is

$$\beta_q'(0) = -\frac{1}{(1-q)(2-q)}, \qquad \partial\beta_q(0) = (-\infty, \beta_q'(0)].$$

Therefore

$$\eta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) + \frac{1}{2-q} \leq -\frac{1}{1-q} + \frac{1}{2-q} = -\frac{1}{(1-q)(2-q)} = \beta_q'(0),$$

and again

$$\eta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) + \frac{1}{2-q} \in \partial\beta_q\left(\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\right).$$

Thus, by the subgradient inequality,

$$\beta_q(\pi(\boldsymbol{\theta})) \geq \beta_q(\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})) + \left(\eta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) + \frac{1}{2-q}\right)\left(\pi(\boldsymbol{\theta}) - \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\right)$$

for h-a.e. $\boldsymbol{\theta}$. Integrating we get

$$\begin{aligned}\Omega_{2-q}(\pi) - \Omega_{2-q}(\pi_{\boldsymbol{\lambda}}) &\geq \int_\Theta \left(\eta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) + \frac{1}{2-q}\right)\left(\pi(\boldsymbol{\theta}) - \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\right) dh(\boldsymbol{\theta}) \\ &= \int_\Theta \eta_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\left(\pi(\boldsymbol{\theta}) - \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\right) dh(\boldsymbol{\theta}),\end{aligned}$$

because both π and $\pi_{\boldsymbol{\lambda}}$ integrate to 1.

Expand $\eta_{\boldsymbol{\lambda}}$:

$$\begin{aligned}&\int_\Theta \eta_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\left(\pi(\boldsymbol{\theta}) - \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\right) dh(\boldsymbol{\theta}) \\ &= \sum_{i=1}^{P} \lambda_i \left(\int_\Theta T_i(\boldsymbol{\theta})\pi(\boldsymbol{\theta}) dh(\boldsymbol{\theta}) - \int_\Theta T_i(\boldsymbol{\theta})\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) dh(\boldsymbol{\theta})\right) - A_{2-q}(\boldsymbol{\lambda})\left(\int_\Theta \pi(\boldsymbol{\theta}) dh(\boldsymbol{\theta}) - \int_\Theta \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) dh(\boldsymbol{\theta})\right) = 0,\end{aligned}$$

since π and $\pi_{\boldsymbol{\lambda}}$ have the same prescribed moments and both are densities. Hence

$$\Omega_{2-q}(\pi) \geq \Omega_{2-q}(\pi_{\boldsymbol{\lambda}}).$$

When $q < 2$,

$$\Omega_{2-q}(\pi) = -\frac{1}{2-q}\mathcal{S}_{2-q}(\pi),$$

and minimizing Ω_{2-q} is equivalent to maximizing $\mathcal{S}_{2-q}$. □

This certifies optimality of a feasible q-exponential candidate. What we have not proven is the existence of $\boldsymbol{\lambda}$ or of a feasible $\pi_{\boldsymbol{\lambda}}$.

Remark. In general we assume

$$q < \frac{D+4}{D+2}, \qquad q \neq 1.$$

Then $q < 1$ yields bounded-support q-Gaussians, while

$$1 < q < \frac{D+4}{D+2}$$

yields Student-t q-Gaussians with

$$\nu = \frac{2}{q-1} - D > 2.$$

Under the parameter range imposed by finite covariance, $q = 2$ is excluded.

B.3 Existence

Our abstract construction in Sec. 3.2 is stated on the canonical domain

$$\Lambda_\Omega = \{\boldsymbol{\lambda} \in \mathbb{R}^P : \Omega^*(\eta_{\boldsymbol{\lambda}}) \text{ is attained by a unique } \pi_{\boldsymbol{\lambda}} \in \mathcal{P}\}.$$

This conditional formulation is deliberate. A general existence theorem for arbitrary Ω would require additional infinite-dimensional assumptions on the support, the moment functions, and the regularizer. For the Shannon case, sufficient conditions of this kind are available in Ishwar and Moulin [2005]. For the Tsallis families used in Sec. F, we do not need such a general result because the optimizer is explicit: the bounded-support case is compactly supported, while the heavy-tailed case reduces to a Student-t family whose normalization and moment conditions can be checked directly from the closed form in App. D.

C Proof of Thm. 1

To generalize Khan and Rue [2023] we derive a canonical gradient descent for optimizing Eq. 14. We use $\bar{\ell}(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - \eta_{\boldsymbol{\lambda}_0}(\boldsymbol{\theta})$. By Prop. 1, the finite-dimensional potential Ψ^* characterizes the geometry of the canonical parameter space. Thus, we define the generalized Fisher matrix,

$$\mathbf{F}(\boldsymbol{\lambda}) = \nabla^2_{\boldsymbol{\lambda}} \Psi^*(\boldsymbol{\lambda}) = \nabla^2_{\boldsymbol{\lambda}} \Omega^*(\eta_{\boldsymbol{\lambda}}) = \frac{\partial \boldsymbol{\mu}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}. \tag{19}$$

and use the chain rule $\mathbf{F}(\boldsymbol{\lambda})^{-1}\nabla_{\boldsymbol{\lambda}} = \mathbf{F}(\boldsymbol{\lambda})^{-1}\mathbf{F}(\boldsymbol{\lambda})\nabla_{\boldsymbol{\mu}} = \nabla_{\boldsymbol{\mu}}$, to obtain a generalized canonical gradient update of the same form as that of Khan and Rue [2023]. Whenever $\mathbf{F}(\boldsymbol{\lambda})$ exists and is invertible, the Fenchel-Young BLR is

$$\boldsymbol{\lambda}_{t+1} \leftarrow \boldsymbol{\lambda}_t - \rho_t \mathbf{F}(\boldsymbol{\lambda}_t)^{-1} \nabla_{\boldsymbol{\lambda}} \left(\mathop{\mathbb{E}}_{\pi_{\boldsymbol{\lambda}}} [\bar{\ell}(\boldsymbol{\theta})] + \Omega(\pi_{\boldsymbol{\lambda}}) \right) \Big|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_t} = (1-\rho_t)\boldsymbol{\lambda}_t - \rho_t \nabla_{\boldsymbol{\mu}} \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\bar{\ell}(\boldsymbol{\theta})]. \tag{20}$$

The equality follows from the fact that $\nabla_{\boldsymbol{\mu}} \Omega(\pi_{\boldsymbol{\lambda}}) = \nabla_{\boldsymbol{\mu}} \Psi(\boldsymbol{\mu}) = \boldsymbol{\lambda}$ (Sec. 3.2), which results in the $(1-\rho_t)$ factor. This characterizes the stationary condition of the generalized variational objective. At any fixed point $\boldsymbol{\lambda}^*$ of the canonical update $\boldsymbol{\lambda}^* = -\nabla_{\boldsymbol{\mu}} \mathbb{E}_{\pi_{\boldsymbol{\lambda}^*}}[\bar{\ell}(\boldsymbol{\theta})]$.

Proof of Thm. 1. By the definition of the Fenchel-Young discrepancy,

$$\mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\ell(\boldsymbol{\theta})] + \mathbb{D}_{\Omega}(\eta_{\boldsymbol{\lambda}_0}, \pi_{\boldsymbol{\lambda}}) = \Omega^*(\eta_{\boldsymbol{\lambda}_0}) + \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\bar{\ell}(\boldsymbol{\theta})] + \Omega(\pi_{\boldsymbol{\lambda}}). \tag{21}$$

By Prop. 1, $\Omega^*(\eta_{\boldsymbol{\lambda}_0}) = \Psi^*(\boldsymbol{\lambda}_0)$, which is constant with respect to $\boldsymbol{\lambda}$. By Prop. 2, each $\pi_{\boldsymbol{\lambda}}$ solves the maxent problem at moment vector $\boldsymbol{\mu}(\boldsymbol{\lambda})$. Hence $\Omega(\pi_{\boldsymbol{\lambda}}) = \Psi(\boldsymbol{\mu}(\boldsymbol{\lambda}))$. Therefore, up to the additive constant $\Psi^*(\boldsymbol{\lambda}_0)$,

$$\mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\ell(\boldsymbol{\theta})] + \mathbb{D}_{\Omega}(\eta_{\boldsymbol{\lambda}_0}, \pi_{\boldsymbol{\lambda}}) = \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\bar{\ell}(\boldsymbol{\theta})] + \Psi(\boldsymbol{\mu}(\boldsymbol{\lambda})).$$

Taking the gradient w.r.t. $\boldsymbol{\lambda}$ and applying the chain rule

$$\begin{aligned} \nabla_{\boldsymbol{\lambda}} \left[\mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\ell(\boldsymbol{\theta})] + \mathbb{D}_{\Omega}(\eta_{\boldsymbol{\lambda}_0}, \pi_{\boldsymbol{\lambda}}) \right] &= \frac{\partial \boldsymbol{\mu}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} \left[\nabla_{\boldsymbol{\mu}} \mathbb{E}_{\pi_{\boldsymbol{\lambda}(\boldsymbol{\mu})}}[\bar{\ell}(\boldsymbol{\theta})] + \nabla_{\boldsymbol{\mu}} \Psi(\boldsymbol{\mu}) \right]_{\boldsymbol{\mu}=\boldsymbol{\mu}(\boldsymbol{\lambda})} \\ &= \mathbf{F}(\boldsymbol{\lambda}) \left[\nabla_{\boldsymbol{\mu}} \mathbb{E}_{\pi_{\boldsymbol{\lambda}(\boldsymbol{\mu})}}[\bar{\ell}(\boldsymbol{\theta})] + \nabla_{\boldsymbol{\mu}} \Psi(\boldsymbol{\mu}) \right]_{\boldsymbol{\mu}=\boldsymbol{\mu}(\boldsymbol{\lambda})}. \end{aligned} \tag{22}$$

Using the dual-coordinate relation $\nabla_{\boldsymbol{\mu}} \Psi(\boldsymbol{\mu}(\boldsymbol{\lambda})) = \boldsymbol{\lambda}$ and multiplying by $\mathbf{F}(\boldsymbol{\lambda})^{-1}$ we get

$$\mathbf{F}(\boldsymbol{\lambda})^{-1} \nabla_{\boldsymbol{\lambda}} \left[\mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\ell(\boldsymbol{\theta})] + \mathbb{D}_{\Omega}(\eta_{\boldsymbol{\lambda}_0}, \pi_{\boldsymbol{\lambda}}) \right] = \nabla_{\boldsymbol{\mu}} \mathbb{E}_{\pi_{\boldsymbol{\lambda}(\boldsymbol{\mu})}}[\bar{\ell}(\boldsymbol{\theta})] \Big|_{\boldsymbol{\mu}=\boldsymbol{\mu}(\boldsymbol{\lambda})} + \boldsymbol{\lambda}.$$

This proves Eq. 15. Substituting this identity into the canonical-gradient step gives

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t - \rho_t \left[\nabla_{\boldsymbol{\mu}} \mathbb{E}_{\pi_{\boldsymbol{\lambda}(\boldsymbol{\mu})}}[\bar{\ell}(\boldsymbol{\theta})] \Big|_{\boldsymbol{\mu}=\boldsymbol{\mu}(\boldsymbol{\lambda}_t)} + \boldsymbol{\lambda}_t \right],$$

which rearranges to

$$\boldsymbol{\lambda}_{t+1} = (1-\rho_t)\boldsymbol{\lambda}_t - \rho_t \nabla_{\boldsymbol{\mu}} \mathbb{E}_{\pi_{\boldsymbol{\lambda}(\boldsymbol{\mu})}}[\bar{\ell}(\boldsymbol{\theta})] \Big|_{\boldsymbol{\mu}=\boldsymbol{\mu}(\boldsymbol{\lambda}_t)}.$$

The fixed-point condition follows by setting $\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t = \boldsymbol{\lambda}^*$ with $\rho_t > 0$. □

D q-Gaussians: q-exponential Family Form to Location-Scale

D.1 Canonical q-exponential form

We start from the q-exponential family canonical parametrization of q-Gaussians (Eq. 81). To derive Alg. 1 and Alg. 2 we use the following sufficient statistics and canonical parameters, which along with the deformation index q determine the q-Gaussian density

$$\mathbf{T}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\theta} \\ \boldsymbol{\theta}\boldsymbol{\theta}^\top \end{pmatrix}, \qquad \boldsymbol{\lambda} = \begin{pmatrix} \mathbf{S}\mathbf{m} \\ -\frac{1}{2}\mathbf{S} \end{pmatrix}. \tag{23}$$

Both bounded-support ($q < 1$), and heavy-tailed densities ($q > 1$) can be written with subtractive normalization in q-exponential form

$$\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \exp_q\left(\boldsymbol{\lambda}^\top \mathbf{T}(\boldsymbol{\theta}) - A_{2-q}(\boldsymbol{\lambda})\right), \qquad \boldsymbol{\lambda}^\top \mathbf{T}(\boldsymbol{\theta}) = -\tfrac{1}{2}(\boldsymbol{\theta}-\mathbf{m})^\top \mathbf{S}(\boldsymbol{\theta}-\mathbf{m}) + \tfrac{1}{2}\mathbf{m}^\top \mathbf{S}\mathbf{m}, \tag{24}$$

with constants absorbed into the deformed cumulant A_{2-q}.

When $\boldsymbol{q < 1}$

$$A_{2-q}(\boldsymbol{\lambda}) = \frac{1}{2}\mathbf{m}^\top \mathbf{S}\mathbf{m} - \frac{1}{2}\left(R(q,D)^2 |\mathbf{S}|^{1/\gamma} - \frac{2}{1-q}\right) \tag{25}$$

with constant $\gamma = D + \frac{2}{1-q}$, and where the max support radius R has a closed form that we derive below.

When $\boldsymbol{q > 1}$

$$A_{2-q}(\boldsymbol{\lambda}) = \frac{1}{2}\mathbf{m}^\top \mathbf{S}\mathbf{m} + \frac{1}{q-1}\left(|\mathbf{S}|^{-1/\nu}\frac{\nu}{\nu+D}\kappa_t(\nu, D) - 1\right) \tag{26}$$

where $\nu = \frac{2}{q-1} - D$ is the Student-t parameter as we show in more detail below where we also define constant κ_t.

D.2 Radial laws and location-scale form

To derive the closed forms of the covariances of $\boldsymbol{\theta}$ that we use to derive the Newton-like updates in App. F.1 for the different ranges of q we need the following polar coordinate stochastic reparametrization. For parameter dimension D we have

$$r^2(\boldsymbol{\theta}) \;=\; (\boldsymbol{\theta}-\mathbf{m})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\mathbf{m}), \qquad \boldsymbol{\theta} = \mathbf{m} + \boldsymbol{\Sigma}^{1/2}\mathbf{z}, \quad \mathbf{z} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\theta}-\mathbf{m}) = r\mathbf{u}, \tag{27}$$

where $\mathbf{u} \sim \mathrm{Unif}(S^{D-1})$ on the unit sphere surface $S^{D-1} = \{\mathbf{u} : \mathbf{u}^\top \mathbf{u} = 1\}$, and $r \geq 0$ is independent of $\mathbf{u}$.

When $\boldsymbol{q < 1}$, we use the determinant-rescaled elliptical scale matrix $\boldsymbol{\Sigma} = |\mathbf{S}|^{1/\gamma}\mathbf{S}^{-1}$ to write $\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ in elliptical Pearson Type II, location-scale form, which is obtained from Eq. 24 and Eq. 25 by the definition of $\exp_q$ (Eq. 82) and some algebraic manipulation

$$\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = |\boldsymbol{\Sigma}|^{-1/2}\left[\frac{1-q}{2}\left(R^2 - r^2(\boldsymbol{\theta})\right)\right]_+^{\frac{1}{1-q}}, \qquad \mathrm{supp}(\pi_{\boldsymbol{\lambda}}) = \{\boldsymbol{\theta} : r^2(\boldsymbol{\theta}) < R^2\}. \tag{28}$$

Evaluating the integral $\int_{\mathbb{R}^D}\left[\frac{1-q}{2}\left(R^2 - \|\mathbf{z}\|^2\right)\right]_+^{\frac{1}{1-q}} d\mathbf{z} = 1$ in terms of the polar coordinate reparametrization from Eq. 27 yields the closed form of the max support radius $R(q, D)$

$$R(q,D) = \left[\frac{\Gamma\left(\frac{D}{2} + \frac{2-q}{1-q}\right)}{\pi^{D/2}\Gamma\left(\frac{2-q}{1-q}\right)}\left(\frac{2}{1-q}\right)^{\frac{1}{1-q}}\right]^{\frac{1-q}{2+D(1-q)}}, \tag{29}$$

where Γ is the Gamma function.

The squared "radial fraction" follows a Beta distribution:

$$\frac{r^2}{R^2(q,D)} \sim \mathrm{Beta}\left(\frac{D}{2}, \frac{2-q}{1-q}\right), \qquad r \in [0, R(q,D)]. \tag{30}$$

Hence **sampling** can be implemented as $\mathbf{u} \sim \mathrm{Unif}(S^{D-1})$, $\frac{r^2}{R^2(q,D)} \sim \mathrm{Beta}(\frac{D}{2}, \frac{2-q}{1-q})$, $\mathbf{z} \leftarrow r\mathbf{u}$, $\boldsymbol{\theta} \leftarrow \mathbf{m} + \boldsymbol{\Sigma}^{1/2}\mathbf{z}$.

When $\boldsymbol{q > 1}$, $\nu = \frac{2}{q-1} - D$ and $\boldsymbol{\Sigma} = \kappa_t |\mathbf{S}|^{-1/\nu}\mathbf{S}^{-1}$ where

$$\kappa_t(\nu, D) = \frac{\nu + D}{\nu}\left[(\nu+D)^{D/2}\pi^{D/2}\frac{\Gamma(\nu/2)}{\Gamma((\nu+D)/2)}\right]^{2/\nu}. \tag{31}$$

From Eq. 24 and Eq. 26 by the definition of $\exp_q$ (Eq. 82) and some algebraic manipulation we obtain the elliptical Pearson Type VII form of the density

$$\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \left(|\mathbf{S}|^{-1/\nu}\frac{\nu}{\nu+D}\kappa_t + \frac{q-1}{2}(\boldsymbol{\theta}-\mathbf{m})^\top\mathbf{S}(\boldsymbol{\theta}-\mathbf{m})\right)^{-\frac{1}{q-1}} \propto \left(1+\frac{1}{\nu}r^2(\boldsymbol{\theta})\right)^{-(\nu+D)/2}. \tag{32}$$

Finally, we can bring the density in Student-t form:

$$\mathcal{T}_\nu(\boldsymbol{\theta} \mid \mathbf{m}, \boldsymbol{\Sigma}) = \frac{1}{Z_\nu|\boldsymbol{\Sigma}|^{1/2}}\left(1+\tfrac{1}{\nu}r^2(\boldsymbol{\theta})\right)^{-(\nu+D)/2}, \qquad Z_\nu = \frac{\Gamma(\nu/2)(\nu\pi)^{D/2}}{\Gamma((\nu+D)/2)}, \tag{33}$$

since with Σ as defined above, Eq. 32 is proportional to the standard Student-t kernel.

In standardized coordinates the Pearson Type VII density is:

$$\pi(\mathbf{z}) = \frac{1}{Z_\nu}\left(1+\tfrac{1}{\nu}\|\mathbf{z}\|^2\right)^{-(\nu+D)/2}, \qquad \mathbf{z}\in\mathbb{R}^D. \tag{34}$$

Using again $\mathbf{z} = r\mathbf{u}$ with Jacobian $d\mathbf{z} = r^{D-1}drd\mathbf{u}$ and integrating out the angular part $\mathbf{u}$, the induced density of the radius r is proportional to

$$r^{D-1}\left(1+\tfrac{1}{\nu}r^2\right)^{-(\nu+D)/2}, \qquad r\in[0,\infty). \tag{35}$$

Equivalently, by the change of variables $r \mapsto r^2$ (i.e. $dr = \frac{1}{2}(r^2)^{-1/2}d(r^2)$), the induced density of r^2 is proportional to

$$(r^2)^{\frac{D}{2}-1}\left(1+\tfrac{1}{\nu}r^2\right)^{-(\nu+D)/2}, \qquad r^2\in(0,\infty), \tag{36}$$

which is a Beta-prime kernel. Hence,

$$\frac{r^2}{\nu} \sim \text{BetaPrime}\left(\frac{D}{2},\frac{\nu}{2}\right). \tag{37}$$

A standard equivalent representation is the chi-square ratio or Fisher-F form:

$$r^2 \overset{d}{=} \nu\frac{\chi^2_D}{\chi^2_\nu} = DF_{D,\nu}, \qquad r\in[0,\infty), \tag{38}$$

with independent chi-squares in the first expression.

We can **sample** $\mathbf{u}\sim\text{Unif}(S^{D-1})$, $U\sim\chi^2_D$, $V\sim\chi^2_\nu$ independently, set $r\leftarrow\sqrt{\nu\,U/V}$, $\mathbf{z}\leftarrow r\,\mathbf{u}$, and finally $\boldsymbol{\theta}\leftarrow\mathbf{m}+\boldsymbol{\Sigma}^{1/2}\mathbf{z}$. Equivalently, sample $F\sim F_{D,\nu}$ and set $r\leftarrow\sqrt{DF}$.

D.3 Expectation parameters

To derive the general mixed Bregman form of the BLR we use linear moments under $\pi_{\boldsymbol{\lambda}}$:

$$\boldsymbol{\mu} = \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\boldsymbol{\theta}] = \mathbf{m}, \qquad \mathbf{M} = \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\boldsymbol{\theta}\boldsymbol{\theta}^\top] = \mathbf{m}\mathbf{m}^\top + \text{Var}_{\pi_{\boldsymbol{\lambda}}}(\boldsymbol{\theta}). \tag{39}$$

For any elliptical $\pi_{\boldsymbol{\lambda}}$ with stochastic representation $\boldsymbol{\theta} = \mathbf{m}+\boldsymbol{\Sigma}^{1/2}(r\mathbf{u})$, using $\boldsymbol{\theta}-\boldsymbol{\mu} = \boldsymbol{\Sigma}^{1/2}(r\mathbf{u})$

$$\text{Var}_{\pi_{\boldsymbol{\lambda}}}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}^{1/2}\mathbb{E}[r^2]\mathbb{E}[\mathbf{u}\mathbf{u}^\top]\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}^{1/2}\mathbb{E}[r^2]\frac{1}{D}\mathbf{I}\boldsymbol{\Sigma}^{1/2} = \frac{\mathbb{E}[r^2]}{D}\boldsymbol{\Sigma}. \tag{40}$$

In particular, from Eq. 40 and the radial laws Eq. 30 and Eq. 38:

$$q<1:\ \text{Var}_{\pi_{\boldsymbol{\lambda}}}(\boldsymbol{\theta}) = \kappa(q,D)\boldsymbol{\Sigma}, \qquad q>1:\ \text{Var}_{\pi_{\boldsymbol{\lambda}}}(\boldsymbol{\theta}) = \frac{\nu}{\nu-2}\boldsymbol{\Sigma}\quad(\nu>2) \tag{41}$$

with $\kappa = \frac{R^2}{D+\frac{2(2-q)}{1-q}}$ as in Alg. 1.

Escort reweightings. The Price- and Stein-type identities that can be derived for q-Gaussians naturally introduce a $(2-q)$-escort (see App. E for more details on related bibliography, and a proof of the heavy-tailed Price-type identity):

$$\tilde{\pi}^{2-q}(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) w_{2-q}(\boldsymbol{\theta}), \tag{42}$$

with

$$q<1: \quad w_{2-q}(\boldsymbol{\theta}) = R^2(q,D) - r^2(\boldsymbol{\theta}), \qquad q>1: \quad w_{2-q}(\boldsymbol{\theta}) = 1 + \frac{1}{\nu} r^2(\boldsymbol{\theta}). \tag{43}$$

D.4 FY Mixed-space Bregman between two heavy-tailed q-Gaussians

Analogously to the expression in [Martins et al., 2022, Prop. 18] we derive the FY mixed-space Bregman divergence for the $q>1$ case. For $\mathbf{T}(\boldsymbol{\theta}) = (\boldsymbol{\theta}, \boldsymbol{\theta}\boldsymbol{\theta}^\top)$ and $\boldsymbol{\lambda} = (\mathbf{S}\mathbf{m}, -\frac{1}{2}\mathbf{S})$, we consider the quadratic form of a potential $\eta_0(\boldsymbol{\theta}) = \boldsymbol{\lambda}_0^\top \mathbf{T}(\boldsymbol{\theta})$ in the same linear family $\boldsymbol{\lambda}_0 = \begin{pmatrix} \mathbf{S}_0\mathbf{m}_0 \\ -\frac{1}{2}\mathbf{S}_0 \end{pmatrix}$. Up to an additive constant independent of $\boldsymbol{\theta}$ $\eta_0(\boldsymbol{\theta}) = -\frac{1}{2}(\boldsymbol{\theta}-\mathbf{m}_0)^\top \mathbf{S}_0 (\boldsymbol{\theta}-\mathbf{m}_0)$. Let π_0 denote the q-Gaussian induced by this potential, i.e.

$$\pi_0(\boldsymbol{\theta}) = \mathcal{T}_\nu(\boldsymbol{\theta} \mid \mathbf{m}_0, \boldsymbol{\Sigma}_0), \qquad \boldsymbol{\Sigma}_0 = \kappa_t |\mathbf{S}_0|^{-1/\nu} \mathbf{S}_0^{-1}. \tag{44}$$

The FY mixed-space Bregman divergence induced by Ω_{2-q} is

$$\mathrm{D}_{\Omega_{2-q}}(\eta_0, \pi) = \Omega_{2-q}(\pi) - \Omega_{2-q}(\pi_0) - \boldsymbol{\lambda}_0^\top \left(\boldsymbol{\mu}(\pi) - \boldsymbol{\mu}(\pi_0)\right), \tag{45}$$

where $\boldsymbol{\mu}(\pi) = (\mathbb{E}_\pi[\boldsymbol{\theta}], \mathbb{E}_\pi[\boldsymbol{\theta}\boldsymbol{\theta}^\top])$. Using $\mathbb{E}_\pi[\boldsymbol{\theta}] = \mathbf{m}$ and $\mathbb{E}_\pi[\boldsymbol{\theta}\boldsymbol{\theta}^\top] = \mathbf{m}\mathbf{m}^\top + \mathrm{Var}_\pi(\boldsymbol{\theta})$, we get

$$\begin{aligned} -\boldsymbol{\lambda}_0^\top (\boldsymbol{\mu}(\pi) - \boldsymbol{\mu}(\pi_0)) &= \frac{1}{2}(\mathbf{m}-\mathbf{m}_0)^\top \mathbf{S}_0 (\mathbf{m}-\mathbf{m}_0) \\ &\quad + \frac{1}{2}\mathrm{Tr}\left(\mathbf{S}_0 \left(\mathrm{Var}_\pi(\boldsymbol{\theta}) - \mathrm{Var}_{\pi_0}(\boldsymbol{\theta})\right)\right). \end{aligned} \tag{46}$$

For Student-t with $\nu>2$, $\mathrm{Var}_\pi(\boldsymbol{\theta}) = \frac{\nu}{\nu-2}\boldsymbol{\Sigma}$, hence

$$\frac{1}{2}\mathrm{Tr}\left(\mathbf{S}_0 \left(\mathrm{Var}_\pi(\boldsymbol{\theta}) - \mathrm{Var}_{\pi_0}(\boldsymbol{\theta})\right)\right) = \frac{\kappa_t \nu}{2(\nu-2)} \left(|\mathbf{S}|^{-1/\nu} \mathrm{Tr}(\mathbf{S}_0 \mathbf{S}^{-1}) - D|\mathbf{S}_0|^{-1/\nu} \right). \tag{47}$$

Using the deformed cumulant expression in Eq. 26 and $\Omega_{2-q}(\pi) = \frac{1}{2-q}(\boldsymbol{\lambda}^\top \boldsymbol{\mu} - A_{2-q}(\boldsymbol{\lambda}))$, we obtain (up to a constant depending only on (ν, D))

$$\Omega_{2-q}(\pi) = \text{const.}(\nu, D) - \frac{\kappa_t \nu(\nu+D)}{2(\nu-2)} |\mathbf{S}|^{-1/\nu}, \tag{48}$$

and the difference simplifies to

$$\Omega_{2-q}(\pi) - \Omega_{2-q}(\pi_0) = -\frac{\kappa_t \nu(\nu+D)}{2(\nu-2)} \left(|\mathbf{S}|^{-1/\nu} - |\mathbf{S}_0|^{-1/\nu} \right). \tag{49}$$

Putting everything together we get the heavy-tail, $q>1$ FY mixed-space Bregman divergence

$$\begin{aligned} \mathrm{D}_{\Omega_{2-q}}(\eta_0, \pi) &= \frac{1}{2}(\mathbf{m}-\mathbf{m}_0)^\top \mathbf{S}_0 (\mathbf{m}-\mathbf{m}_0) \\ &\quad + \frac{\kappa_t \nu}{2(\nu-2)} \left[|\mathbf{S}|^{-1/\nu} \left(\mathrm{Tr}(\mathbf{S}_0 \mathbf{S}^{-1}) - (\nu+D) \right) + \nu |\mathbf{S}_0|^{-1/\nu} \right]. \end{aligned} \tag{50}$$

E Bonnet-, Price-, and Stein-type identities for q-Gaussians

In Alg. 1 and Alg. 2 we replace the Gaussian samples of VON with elliptical q-Gaussian samples, using generalized Bonnet and Price identities related to appropriate elliptical Stein-type identities. The resulting q-VON algorithms retain VON's form, differing by factors depending only on (q, D) and closed-form escort reweighting derived from the dual; as $q \to 1$ we recover standard VON.

By generalizing the VON sampling step from a Gaussian to a q-Gaussian distribution we obtain a more general class of scalable q-VON algorithms, which extend the BLR to optimize over a much wider class of distributions. We focus on variants of VON involving deformed exponential families with bounded, parameter-dependent support ($q < 1$), and with heavy tails ($q > 1$). Elliptical families include q-Gaussians, which are encompassed under Pearson Type II (bounded support) and Type VII (heavy-tailed) distributions for appropriate ranges of q, and provide bounded-support and heavy-tailed behaviors with tractable geometry and reparameterizations. Further, they admit Stein-type identities which lead to generalized Bonnet and Price theorems [Landsman and Nešlehová, 2008, Landsman et al., 2015, Matsuzoe and Ohara, 2011]. The references concerning the heavy-tailed cases prove Stein-type identities but do not apply them to obtain Bonnet- and Price-type theorems, which we prove below for reference. Related proofs can be found in Khan and Lin [2017], Lukashchuk et al. [2026]. A bounded-support q-Gaussian Stein identity is derived and used to prove Bonnet- and Price-type identities in Sklaviadis et al. [2026].

Here we review the gradient identities used to derive the q-VON updates (App. F.1, Alg. 1 and Alg. 2). Throughout, $\boldsymbol{\theta} \in \mathbb{R}^D$ is the random variable and $\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ is the variational posterior with canonical parameters $\boldsymbol{\lambda} = \left(\mathbf{S}\mathbf{m}, -\frac{1}{2}\mathbf{S}\right)$. We write the same distribution in elliptical location-scale form $\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = q_q(\boldsymbol{\theta} \mid \mathbf{m}, \boldsymbol{\Sigma})$, with $r^2(\boldsymbol{\theta}) = (\boldsymbol{\theta} - \mathbf{m})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \mathbf{m})$.

E.1 Bonnet-type identities

For both $q < 1$ and $q > 1$, for differentiable f with suitable integrability, the gradient with respect to the location parameter satisfies the Bonnet identity

$$\nabla_{\mathbf{m}} \mathbb{E}_{\pi_{\boldsymbol{\lambda}}} \left[f(\boldsymbol{\theta}) \right] = \mathbb{E}_{\pi_{\boldsymbol{\lambda}}} \left[\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \right]. \tag{51}$$

E.1.1 Proof of the $q > 1$ or Student-t Bonnet-type identity

For a proof of the $q < 1$ case see Sklaviadis et al. [2026].

Let $\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \mathcal{T}_\nu(\boldsymbol{\theta} \mid \mathbf{m}, \boldsymbol{\Sigma})$ with $\nu > 2$ and support $\mathbb{R}^D$,

$$\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \frac{1}{Z_\nu |\boldsymbol{\Sigma}|^{1/2}} \left(1 + \tfrac{1}{\nu} r^2(\boldsymbol{\theta})\right)^{-(\nu+D)/2}. \tag{52}$$

Assume $f : \mathbb{R}^D \to \mathbb{R}$ is differentiable and integrable under $\pi_{\boldsymbol{\lambda}}$, and that boundary terms vanish (e.g. f and $\nabla_{\boldsymbol{\theta}} f$ are bounded; it suffices that $\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ decays polynomially).

We show Eq. 51 holds componentwise. Fix $i \in \{1, \dots, D\}$. By dominated convergence we may differentiate under the integral:

$$\frac{\partial}{\partial \mathbf{m}_i} \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[f(\boldsymbol{\theta})] = \int_{\mathbb{R}^D} f(\boldsymbol{\theta}) \frac{\partial}{\partial \mathbf{m}_i} \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) \, d\boldsymbol{\theta}. \tag{53}$$

Since $\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ depends on $(\boldsymbol{\theta}, \mathbf{m})$ only through $(\boldsymbol{\theta} - \mathbf{m})$, we have the identity

$$\frac{\partial}{\partial \mathbf{m}_i} \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = -\frac{\partial}{\partial \boldsymbol{\theta}_i} \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}). \tag{54}$$

Therefore,

$$\frac{\partial}{\partial \mathbf{m}_i} \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[f(\boldsymbol{\theta})] = -\int f(\boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}_i} \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) d\boldsymbol{\theta} = -\int \frac{\partial}{\partial \boldsymbol{\theta}_i} \left(\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) f(\boldsymbol{\theta})\right) d\boldsymbol{\theta} + \int \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) \frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i} d\boldsymbol{\theta}. \tag{55}$$

The total-derivative term vanishes by the boundary assumption, and thus

$$\frac{\partial}{\partial \mathbf{m}_i} \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[f(\boldsymbol{\theta})] = \mathbb{E}_{\pi_{\boldsymbol{\lambda}}} \left[\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i} \right]. \tag{56}$$

Stacking coordinates gives the vector form Eq. 51: $\nabla_{\mathbf{m}} \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[f(\boldsymbol{\theta})] = \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})]$.

E.2 Price-type identities

Let f be twice differentiable with $\nabla_{\boldsymbol{\theta}} f$ bounded and $\nabla^2_{\boldsymbol{\theta}} f$ integrable. Then the gradient with respect to the scale parameter admits a generalized Price identity of the form

$$\nabla_{\boldsymbol{\Sigma}} \mathbb{E}_{\pi_{\boldsymbol{\lambda}}} \left[f(\boldsymbol{\theta}) \right] = \frac{1}{2} \mathbb{E}_{\pi_{\boldsymbol{\lambda}}} \left[\tau_q(\boldsymbol{\theta}) \nabla^2_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \right], \tag{57}$$

where the scalar factor $\tau_q(\boldsymbol{\theta})$ depends on the value of q:

$$\tau_q(\boldsymbol{\theta}) = \begin{cases} \dfrac{1-q}{2(2-q)} w_{2-q}, & q < 1, \\ \dfrac{\nu}{\nu + D - 2} w_{2-q} = \dfrac{\nu + r^2(\boldsymbol{\theta})}{\nu + D - 2}, & q > 1, \end{cases} \quad \text{where} \tag{58}$$

$$w_{2-q}(\boldsymbol{\theta}) = \begin{cases} R^2(q, D) - r^2(\boldsymbol{\theta}), & q < 1, \\ 1 + \frac{1}{\nu} r^2(\boldsymbol{\theta}), & q > 1. \end{cases} \tag{59}$$

Escort form. For both ranges of q

$$\tilde{\pi}^{2-q}(\boldsymbol{\theta}) \propto \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) w_{2-q}(\boldsymbol{\theta}). \tag{60}$$

This means that Eq. 57 can be rewritten compactly in a form that resembles the Gaussian Stein identity as

$$\nabla_{\boldsymbol{\Sigma}} \mathbb{E}_{\pi_{\boldsymbol{\lambda}}} \left[f(\boldsymbol{\theta}) \right] = \frac{c_q}{2} \mathbb{E}_{\tilde{\pi}^{2-q}} \left[\nabla^2_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \right], \tag{61}$$

with the constant

$$c_q = \mathbb{E}_{\pi} \tau_q(\boldsymbol{\theta}) = \begin{cases} \kappa(q, D), & q < 1, \\ \dfrac{\nu}{\nu - 2}, & q > 1 \ (\nu > 2). \end{cases} \tag{62}$$

When $q \to 1$ we retrieve the Gaussian identity $\nabla_{\boldsymbol{\Sigma}} \mathbb{E}_{\mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma})}[f] = \frac{1}{2} \mathbb{E}[\nabla^2_{\boldsymbol{\theta}} f]$.

E.2.1 Diagonal Stein estimator

Eq. (61) shows that the gradient with respect to the scale parameter depends on the Hessian of f averaged under the complementary escort $\tilde{\pi}^{2-q}$. In our q-VON algorithms we restrict to diagonal $\boldsymbol{\Sigma}$ and only require the diagonal of $\mathbb{E}_{\tilde{\pi}^{2-q}}[\nabla^2_{\boldsymbol{\theta}} f(\boldsymbol{\theta})]$. To avoid explicit Hessian evaluations, we use a Stein identity to rewrite these escort Hessian diagonal entries in terms of first derivatives evaluated at samples $\boldsymbol{\theta} \sim \pi_{\boldsymbol{\lambda}}$.

Let $f : \mathbb{R}^D \to \mathbb{R}$ be differentiable and integrable under $\pi_{\boldsymbol{\lambda}}$. Stein-type identities take the following form

$$\mathbb{E}_{\pi_{\boldsymbol{\lambda}}} \left[\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \right] = -\mathbb{E}_{\pi_{\boldsymbol{\lambda}}} \left[f(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) \right], \tag{63}$$

provided boundary terms vanish. For $q > 1$ (unbounded support) a Stein-type identity has been proven by Landsman et al. [2015]. For $q < 1$ the support is bounded but $\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ vanishes at the boundary, so a similar identity holds under mild regularity conditions on f as shown in Sklaviadis et al. [2026].

For q-Gaussian densities with scale matrix $\boldsymbol{\Sigma}$ the scores can be written explicitly as

$$\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = -\frac{1}{(2-q)\tau_q(\boldsymbol{\theta})} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta} - \mathbf{m}) \tag{64}$$

Equivalently, and more conveniently for our purposes, these score functions imply an escort covariance form Stein identity:

$$\mathbb{E}_{\pi_{\boldsymbol{\lambda}}} \left[(\boldsymbol{\theta} - \mathbf{m}) f(\boldsymbol{\theta}) \right] = c_q \boldsymbol{\Sigma} \mathbb{E}_{\tilde{\pi}^{2-q}} \left[\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \right]. \tag{65}$$

Algs. 1-2 only require an estimate of the diagonal $\mathbb{E}_{\tilde{\pi}^{2-q}}[\nabla^2_{\boldsymbol{\theta}} f(\boldsymbol{\theta})]$. Assume $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\sigma}^2)$ and define $\widehat{\mathbf{g}} = \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$. Then, componentwise,

$$\mathbb{E}_{\tilde{\pi}^{2-q}} \left[\frac{\partial^2 f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i^2} \right] = \frac{1}{c_q} \mathbb{E}_{\pi_{\boldsymbol{\lambda}}} \left[\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i} \frac{\boldsymbol{\theta}_i - \mathbf{m}_i}{\boldsymbol{\sigma}_i^2} \right], \qquad i = 1, \ldots, D. \tag{66}$$

Thus a single-sample unbiased estimator for the diagonal Hessian under the complementary escort is

$$\widehat{\mathbf{h}} = \frac{1}{c_q}\widehat{\mathbf{g}} \cdot (\boldsymbol{\theta} - \mathbf{m})/\boldsymbol{\sigma}^2, \quad \boldsymbol{\theta} \sim \pi_{\boldsymbol{\lambda}}, \tag{67}$$

which is exactly the form used in Algs. 1 and 2.

E.2.2 Proof of the $q > 1$ Student-t Price-type identity

For a proof of the $q < 1$ Price-type identity see Sklaviadis et al. [2026].

Let $\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \mathcal{T}_\nu(\boldsymbol{\theta} \mid \mathbf{m}, \boldsymbol{\Sigma})$ with $\nu > 2$. Assume f is twice continuously differentiable, integrable under $\pi_{\boldsymbol{\lambda}}$, and that boundary terms vanish for the integrations by parts below (it is sufficient that f and $\nabla_{\boldsymbol{\theta}} f$ are bounded).

We prove the componentwise identity

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_{ij}} \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[f(\boldsymbol{\theta})] = \frac{1}{2}\mathbb{E}_{\pi_{\boldsymbol{\lambda}}}\left[\frac{\nu + r^2(\boldsymbol{\theta})}{\nu + D - 2}\, \frac{\partial^2 f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i\, \partial \boldsymbol{\theta}_j}\right], \qquad i, j \in \{1, \dots, D\}, \tag{68}$$

which is equivalent to Eq. 57 for $q > 1$ since $\frac{\nu+r^2}{\nu+D-2} = \frac{\nu}{\nu+D-2}\left(1 + \frac{1}{\nu}r^2\right)$.

Derivative with respect to $\boldsymbol{\Sigma}$ of the expectation. By dominated convergence,

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_{ij}} \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\Sigma}_{ij}} \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) d\boldsymbol{\theta} = \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}\left[f(\boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\Sigma}_{ij}} \log \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\right]. \tag{69}$$

Writing $\log \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \text{const.} - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{\nu+D}{2}\log\left(1 + \frac{1}{\nu}r^2(\boldsymbol{\theta})\right)$, and using $\frac{\partial}{\partial \boldsymbol{\Sigma}_{ij}} \log|\boldsymbol{\Sigma}| = (\boldsymbol{\Sigma}^{-1})_{ji}$ and $\frac{\partial}{\partial \boldsymbol{\Sigma}_{ij}} r^2(\boldsymbol{\theta}) = -(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \mathbf{m})(\boldsymbol{\theta} - \mathbf{m})^\top \boldsymbol{\Sigma}^{-1})_{ij}$, we obtain

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_{ij}} \log \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = -\frac{1}{2}(\boldsymbol{\Sigma}^{-1})_{ij} + \frac{\nu + D}{2(\nu + r^2(\boldsymbol{\theta}))}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \mathbf{m})(\boldsymbol{\theta} - \mathbf{m})^\top \boldsymbol{\Sigma}^{-1}\right)_{ij}. \tag{70}$$

Define $\mathbf{a}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \mathbf{m})$ so that $r^2(\boldsymbol{\theta}) = (\boldsymbol{\theta} - \mathbf{m})^\top \mathbf{a}(\boldsymbol{\theta})$ and $\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \mathbf{m})(\boldsymbol{\theta} - \mathbf{m})^\top \boldsymbol{\Sigma}^{-1}\right)_{ij} = \mathbf{a}_i(\boldsymbol{\theta})\mathbf{a}_j(\boldsymbol{\theta})$. Then Eq. 70 becomes

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_{ij}} \log \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = -\frac{1}{2}(\boldsymbol{\Sigma}^{-1})_{ij} + \frac{\nu + D}{2(\nu + r^2(\boldsymbol{\theta}))}\mathbf{a}_i(\boldsymbol{\theta})\mathbf{a}_j(\boldsymbol{\theta}). \tag{71}$$

Therefore,

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_{ij}} \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[f(\boldsymbol{\theta})] = \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}\left[f(\boldsymbol{\theta})\left(-\frac{1}{2}(\boldsymbol{\Sigma}^{-1})_{ij} + \frac{\nu + D}{2(\nu + r^2(\boldsymbol{\theta}))}\mathbf{a}_i(\boldsymbol{\theta})\mathbf{a}_j(\boldsymbol{\theta})\right)\right]. \tag{72}$$

Integration by parts to express Eq. 72 in terms of the Hessian. We use the Student-t score:

$$\frac{\partial}{\partial \boldsymbol{\theta}_i} \log \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = -\frac{\nu + D}{\nu + r^2(\boldsymbol{\theta})}\mathbf{a}_i(\boldsymbol{\theta}). \tag{73}$$

An important simplification is

$$\frac{\partial}{\partial \boldsymbol{\theta}_i}\left(\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})(\nu + r^2(\boldsymbol{\theta}))\right) = -(\nu + D - 2)\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\mathbf{a}_i(\boldsymbol{\theta}), \tag{74}$$

which follows from the product rule, Eq. 73, and $\frac{\partial}{\partial \boldsymbol{\theta}_i} r^2(\boldsymbol{\theta}) = 2\,\mathbf{a}_i(\boldsymbol{\theta})$.

Let $w(\boldsymbol{\theta}) = \frac{\nu + r^2(\boldsymbol{\theta})}{\nu + D - 2}$. Then Eq. 74 is equivalent to

$$\frac{\partial}{\partial \boldsymbol{\theta}_i}\left(\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})w(\boldsymbol{\theta})\right) = -\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\mathbf{a}_i(\boldsymbol{\theta}). \tag{75}$$

Now integrate by parts twice:

$$\begin{aligned}\mathbb{E}_{\pi_{\boldsymbol{\lambda}}}\left[w(\boldsymbol{\theta})\frac{\partial^2 f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j}\right] &= \int \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})w(\boldsymbol{\theta})\frac{\partial^2 f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} d\boldsymbol{\theta} = -\int \frac{\partial}{\partial \boldsymbol{\theta}_i}\left(\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})w(\boldsymbol{\theta})\right)\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j} d\boldsymbol{\theta} \\ &\overset{\text{Eq. 75}}{=} \int \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\mathbf{a}_i(\boldsymbol{\theta})\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j} d\boldsymbol{\theta} = -\int \frac{\partial}{\partial \boldsymbol{\theta}_j}\left(\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\mathbf{a}_i(\boldsymbol{\theta})\right) f(\boldsymbol{\theta}) d\boldsymbol{\theta}.\end{aligned} \tag{76}$$

We compute the derivative inside the last integral. Since $\frac{\partial}{\partial \boldsymbol{\theta}_j}\mathbf{a}_i(\boldsymbol{\theta}) = (\boldsymbol{\Sigma}^{-1})_{ij}$ and $\frac{\partial}{\partial \boldsymbol{\theta}_j}\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\frac{\partial}{\partial \boldsymbol{\theta}_j}\log \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$, we get

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}_j}\left(\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\mathbf{a}_i(\boldsymbol{\theta})\right) &= \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})(\boldsymbol{\Sigma}^{-1})_{ij} + \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\mathbf{a}_i(\boldsymbol{\theta})\frac{\partial}{\partial \boldsymbol{\theta}_j}\log \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) \\ &\overset{Eq.\ 73}{=} \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})(\boldsymbol{\Sigma}^{-1})_{ij} - \pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\mathbf{a}_i(\boldsymbol{\theta})\frac{\nu + D}{\nu + r^2(\boldsymbol{\theta})}\mathbf{a}_j(\boldsymbol{\theta}). \end{aligned} \tag{77}$$

Substituting into Eq. 76 yields

$$\mathbb{E}_{\pi_{\boldsymbol{\lambda}}}\left[w(\boldsymbol{\theta})\frac{\partial^2 f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j}\right] = \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}\left[f(\boldsymbol{\theta})\left(-(\boldsymbol{\Sigma}^{-1})_{ij} + \frac{\nu + D}{\nu + r^2(\boldsymbol{\theta})}\mathbf{a}_i(\boldsymbol{\theta})\mathbf{a}_j(\boldsymbol{\theta})\right)\right]. \tag{78}$$

Comparing Eq. 78 with Eq. 72 gives exactly the Price-type identity Eq. 68:

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_{ij}}\mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[f(\boldsymbol{\theta})] = \frac{1}{2}\mathbb{E}_{\pi_{\boldsymbol{\lambda}}}\left[w(\boldsymbol{\theta})\frac{\partial^2 f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j}\right], \tag{79}$$

which completes the proof.

F Tsallis Negentropy and q-VON

As the main example we replace the Shannon negentropy by the Tsallis negative entropy Ω_{2-q}, which is minimized by q-exponential families (see App. B for a proof). When we use quadratic statistics the maxent family becomes a q-Gaussian, which has bounded support for $q < 1$ and heavy tails for $q > 1$. Respectively, these lead to the q-VON and Student-t-VON algorithms. Thus, the quadratic statistic remains the same, while the choice of Ω changes the support and tail behavior.

$$\Omega_{2-q}(\pi) = \begin{cases} \frac{\mathbb{E}_\pi[\pi^{1-q}-1]}{\left((2-q)^2-(2-q)\right)} & \text{if } q \neq 1, \\ \mathbb{E}_\pi[\log \pi] & \text{if } q = 1, \end{cases} \tag{80}$$

Using entropic index $2-q$, and linear expectation constraints, the maxent problem in Eq. 11 is optimized by q-exponential families (see App. B.2 for a proof):[9]

$$\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \exp_q(\boldsymbol{\lambda}^\top \mathbf{T}(\boldsymbol{\theta}) - A_{2-q}(\boldsymbol{\lambda})), \tag{81}$$

where $A_{2-q}(\boldsymbol{\lambda})$ is the deformed cumulant (we give detailed closed form examples in App. D; see also Naudts [2011], Amari and Ohara [2011], Amari [2016], Martins et al. [2022]). The q-exp and q-log functions are defined as

$$\exp_q(u) = \begin{cases} [1 + (1-q)u]_+^{1/(1-q)} & \text{if } q \neq 1 \\ \exp u & \text{if } q = 1 \end{cases}, \qquad \log_q(u) = \begin{cases} \frac{u^{1-q}-1}{1-q} & \text{if } q \neq 1 \\ \log u & \text{if } q = 1 \end{cases}, \tag{82}$$

and exponential families are recovered when $q = 1$.

We will repeatedly use escort distributions associated with a density π. For any order $a > 0$ such that $\int \pi(\boldsymbol{\theta})^a d\boldsymbol{\theta} < \infty$, we define the escort of order a by

$$\tilde{\pi}^a(\boldsymbol{\theta}) = \frac{\pi(\boldsymbol{\theta})^a}{\int \pi(\boldsymbol{\theta})^a d\boldsymbol{\theta}}.$$

In the Tsallis setting below, the orders $a = q$ and $a = 2 - q$ arise naturally in the gradient identities and in the Newton-like updates.

When q-exponential families are derived by maximum Tsallis entropy subject to linear expectation constraints, A_{2-q} is related to Ω and Ω^* as follows [Martins et al., 2022, Prop.10, App. B]:

$$\begin{aligned} \Omega_{2-q}(\pi_{\boldsymbol{\lambda}}) &= \frac{1}{2-q}(\boldsymbol{\lambda}^\top \boldsymbol{\mu}(\boldsymbol{\lambda}) - A_{2-q}(\boldsymbol{\lambda})) \\ \Omega^*_{2-q}(\eta_{\boldsymbol{\lambda}}) &= (1-q)\Omega_{2-q}(\pi_{\boldsymbol{\lambda}}) + A_{2-q}(\boldsymbol{\lambda}) = \frac{1}{2-q}\left((1-q)\boldsymbol{\lambda}^\top \boldsymbol{\mu}(\boldsymbol{\lambda}) + A_{2-q}(\boldsymbol{\lambda})\right), \end{aligned} \tag{83}$$

[9] We assume a constant base measure for simplicity; what follows would work with general base measures.

where $\boldsymbol{\mu}(\boldsymbol{\lambda}) = \mathbb{E}_{\boldsymbol{\theta}\sim\pi_{\boldsymbol{\lambda}}}[\mathbf{T}(\boldsymbol{\theta})] = \nabla_{\boldsymbol{\lambda}}\Omega^*_{2-q}(\eta_{\boldsymbol{\lambda}})$, and

$$\frac{\partial\boldsymbol{\mu}(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}} = \nabla^2_{\boldsymbol{\lambda}}\Omega^*_{2-q}(\eta_{\boldsymbol{\lambda}}) =: \mathbf{F}_{2-q}(\boldsymbol{\lambda}). \tag{84}$$

Eq. 84 is useful in invoking the chain rule to convert gradients into natural or canonical gradients and vice-versa, and the Hessian of Ω^* with respect to the canonical parameters $\mathbf{F}_{2-q}$ recovers the Fisher matrix when $q = 1$.

From Eq. 83 we obtain:

$$\begin{aligned}
\nabla_{\boldsymbol{\mu}}\Omega_{2-q}(\pi_{\boldsymbol{\lambda}}) &= \nabla_{\boldsymbol{\mu}}\mathbb{E}_{\pi_{\boldsymbol{\lambda}}}\left[\frac{1}{2-q}\log_q\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\right] \\
&= \frac{1}{2-q}\nabla_{\boldsymbol{\mu}}\int_{\boldsymbol{\theta}}\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta})[\boldsymbol{\lambda}^\top\mathbf{T}(\boldsymbol{\theta}) - A_{2-q}(\boldsymbol{\lambda})] \\
&= \frac{1}{2-q}\nabla_{\boldsymbol{\mu}}[\boldsymbol{\lambda}^\top\underbrace{\mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\mathbf{T}(\boldsymbol{\theta})]}_{\boldsymbol{\mu}(\boldsymbol{\lambda})} - A_{2-q}(\boldsymbol{\lambda})] \\
&= \frac{1}{2-q}\left[\frac{\partial\boldsymbol{\lambda}}{\partial\boldsymbol{\mu}}\boldsymbol{\mu} + \boldsymbol{\lambda} - \frac{\partial\boldsymbol{\lambda}}{\partial\boldsymbol{\mu}}\frac{\partial A_{2-q}(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}\right] \\
&= \frac{1}{2-q}\left[\boldsymbol{\lambda} + \mathbf{F}^{-1}_{2-q}(\boldsymbol{\lambda})(\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}}_q)\right].
\end{aligned}$$

We define $\widetilde{\boldsymbol{\mu}}_q = \nabla A_{2-q}(\boldsymbol{\lambda}) = \mathbb{E}_{\tilde{\pi}^q}[\mathbf{T}(\boldsymbol{\theta})]$ as the expectation with respect to the escort distribution $\tilde{\pi}^q$ of order q, which is defined as $\tilde{\pi}^q = \pi^q / \int \pi^q d\boldsymbol{\theta}$. Eq. 83 further implies that

$$\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}_q = (1-q)\mathbf{F}_{2-q}\boldsymbol{\lambda}, \tag{85}$$

and substituting above, it follows that when Ω is the Tsallis negentropy

$$\nabla_{\boldsymbol{\mu}}\Omega_{2-q}(\pi_{\boldsymbol{\lambda}}) = \frac{1}{2-q}(\boldsymbol{\lambda} + (1-q)\boldsymbol{\lambda}) = \boldsymbol{\lambda}, \tag{86}$$

which we use below to derive Newton-like generalized BLR algorithms (Alg. 1, Alg. 2).

F.1 q-Variational Online Newton

We now specialize the generalized BLR to the quadratic statistic $\mathbf{T}(\boldsymbol{\theta})^\top = (\boldsymbol{\theta}, \boldsymbol{\theta}\boldsymbol{\theta}^\top)$, $\boldsymbol{\theta} \in \mathbb{R}^D$, which induces a location-scale family of q-Gaussians. The derivation in this subsection proceeds in three steps. First, we write the generic canonical update for this quadratic family in terms of the expectation parameters $(\boldsymbol{\mu}, \mathbf{M})$. Second, we convert gradients with respect to $(\boldsymbol{\mu}, \mathbf{M})$ into gradients with respect to the location-scale parameters $(\mathbf{m}, \boldsymbol{\Sigma})$. Third, we specialize our result to bounded-support $q < 1$ and the heavy-tailed $q > 1$ cases, where Bonnet- and Price-type identities yield the Newton-like updates implemented in Alg. 1 and Alg. 2.

First, we connect the pseudocode to the general canonical update in Eq. 16. In both $q < 1$ and $q > 1$ cases we start from the same canonical update, and what changes is the relation between expectation parameters and location-scale parameters, as well as the corresponding stochastic identities used to evaluate the gradients. In Variational Online Newton (VON) gradients and Hessians are estimated by sampling from a Gaussian variational posterior Khan et al. [2018]. VON-style natural-gradient steps result in fast convergence and competitive test accuracy in large-scale deep learning applications [Shen et al., 2024]. Here, we generalize these updates to bounded-support (Alg. 1) and heavy-tailed (Alg. 2) q-Gaussian posteriors in a unified way. Remarkably, our generalizations require only small changes in the VON pseudocode [Shen et al., 2024, Alg. 1], highlighted in blue.

Using the Tsallis negentropy Ω_{2-q} in Eq. 16 and sufficient statistics $\mathbf{T}(\boldsymbol{\theta})^\top = (\boldsymbol{\theta}, \boldsymbol{\theta}\boldsymbol{\theta}^\top)$ we obtain q-Gaussian families with canonical parameters $\boldsymbol{\lambda}^\top = (\mathbf{S}\mathbf{m}, -\frac{1}{2}\mathbf{S})$. q-Gaussians include both bounded-support and heavy-tailed (Student-t) subfamilies. The sufficient statistics and canonical parameters determine the elliptical geometry (that is, the quadratic form), but the value of q determines the support and tail behavior. We derive the specific forms that Eq. 16 takes for $q < 1$ and $q > 1$, resulting respectively in Alg. 1 and Alg. 2, where in blue we highlight differences from the BLR's

Variational Online Newton form Khan et al. [2018] and IVON [Shen et al., 2024, Alg. 1]. Detailed background on q-Gaussians can be found in App. D.

We begin by writing Eq. 16 more explicitly as

$$\boldsymbol{\lambda}_{t+1} = (1-\rho_t)\boldsymbol{\lambda}_t - \rho_t \begin{bmatrix} \nabla_{\boldsymbol{\mu}} \mathbb{E}_{\pi_{\boldsymbol{\lambda}}} \bar{\ell}(\boldsymbol{\theta}) \\ \nabla_{\mathbf{M}} \mathbb{E}_{\pi_{\boldsymbol{\lambda}}} \bar{\ell}(\boldsymbol{\theta}) \end{bmatrix}. \tag{87}$$

For both $q < 1$ and $q > 1$ the expectation parameters are

$$\begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{M} \end{bmatrix} = \begin{bmatrix} \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\boldsymbol{\theta}] \\ \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[\boldsymbol{\theta}\boldsymbol{\theta}^\top] \end{bmatrix} = \begin{bmatrix} \mathbf{m} \\ \mathrm{Var}_{\pi_{\boldsymbol{\lambda}}}(\boldsymbol{\theta}) + \mathbf{m}\mathbf{m}^\top \end{bmatrix}. \tag{88}$$

Importantly, $\mathrm{Var}(\boldsymbol{\theta})$ has distinct forms for $q < 1$ and $q > 1$, and it is *not* equal to the canonical parameter $\mathbf{S}^{-1}$ or the scale parameter $\boldsymbol{\Sigma}$, unless $q = 1$ in which case $\mathbf{S}^{-1} = \boldsymbol{\Sigma} = \mathrm{Var}(\boldsymbol{\theta})$.

Chain rule. To use Bonnet- and Price-type stochastic gradient identities, we need a chain rule to express gradients with respect to expectation parameters $(\boldsymbol{\mu}, \mathbf{M})$ as gradients with respect to location-scale parameters $\mathbf{m}(\boldsymbol{\mu}), \boldsymbol{\Sigma}(\boldsymbol{\mu}, \mathbf{M})$. For any function $f(\mathbf{m}(\boldsymbol{\mu}), \boldsymbol{\Sigma}(\boldsymbol{\mu}, \mathbf{M}))$, a straightforward generalization of the well known Gaussian chain rule [Opper and Archambeau, 2009, Khan and Lin, 2017] is

$$\nabla_{\boldsymbol{\mu}} f(\mathbf{m}, \boldsymbol{\Sigma}) = \nabla_{\mathbf{m}} f \nabla_{\boldsymbol{\mu}} \mathbf{m} + \nabla_{\boldsymbol{\Sigma}} f \nabla_{\boldsymbol{\mu}} \boldsymbol{\Sigma}, \tag{89}$$

$$\nabla_{\mathbf{M}} f(\mathbf{m}, \boldsymbol{\Sigma}) = \nabla_{\mathbf{m}} f \nabla_{\mathbf{M}} \mathbf{m} + \nabla_{\boldsymbol{\Sigma}} f \nabla_{\mathbf{M}} \boldsymbol{\Sigma}. \tag{90}$$

Bounded support. When $q < 1$, bounded-support q-Gaussians have maximum support radius $R(q, D)$. The max support radius has a closed-form (reviewed in App. D), which avoids dependence on $\mathbf{S}^{-1}$ by expressing the scale matrix as a determinant reweighting $\boldsymbol{\Sigma} = |\mathbf{S}|^{1/\gamma}\mathbf{S}^{-1}$ with constant $\gamma = D + \frac{2}{1-q}$. From Eq. 88 we can show $\mathrm{Var}_{\pi_{\boldsymbol{\lambda}}}(\boldsymbol{\theta}) = \kappa\boldsymbol{\Sigma}$ with constant $\kappa = R^2/(D + 2(2-q)/(1-q))$ (see App. D.3).

From Eq. 89 and $\partial\boldsymbol{\Sigma}/\partial\boldsymbol{\mu} = -2\kappa^{-1}\mathbf{m}$, it follows that

$$\nabla_{\boldsymbol{\mu}} f = \nabla_{\mathbf{m}} f - \frac{2}{\kappa} \nabla_{\boldsymbol{\Sigma}} f \mathbf{m}, \tag{91}$$

and from Eq. 90 $\nabla_{\mathbf{M}}\mathbf{m} = 0$ and $\partial\boldsymbol{\Sigma}/\partial\mathbf{M} = \kappa^{-1}$ that

$$\nabla_{\mathbf{M}} f = \kappa^{-1} \nabla_{\boldsymbol{\Sigma}} f. \tag{92}$$

Setting $f = \bar{\ell}$, Eq. 87 becomes

$$\boldsymbol{\lambda}_{t+1} = (1-\rho_t)\boldsymbol{\lambda}_t - \rho_t \begin{bmatrix} \nabla_{\mathbf{m}} \mathop{\mathbb{E}}_{\boldsymbol{\theta}\sim\pi_{\boldsymbol{\lambda}}} \bar{\ell}(\boldsymbol{\theta}) - 2\kappa^{-1} \left(\nabla_{\boldsymbol{\Sigma}} \mathop{\mathbb{E}}_{\boldsymbol{\theta}\sim\pi_{\boldsymbol{\lambda}}} [\bar{\ell}(\boldsymbol{\theta})] \right) \mathbf{m} \\ \kappa^{-1} \nabla_{\boldsymbol{\Sigma}} \mathop{\mathbb{E}}_{\boldsymbol{\theta}\sim\pi_{\boldsymbol{\lambda}}} [\bar{\ell}(\boldsymbol{\theta})] \end{bmatrix}. \tag{93}$$

Rearranging Eq. 93, and using Price- and Bonnet-type identities for bounded-support q-Gaussians, to push the gradients inside the expectations [Sklaviadis et al., 2026], we obtain the bounded-support Newton-like generalized BLR (Alg. 1):

$$\mathbf{S}_{t+1} = (1-\rho_t)\mathbf{S}_t + \rho_t \frac{1}{\kappa D} \left(\mathbb{E}_{\pi_t}[r^2(\boldsymbol{\theta})] \mathbb{E}_{\tilde{\pi}_t^{2-q}}[\nabla^2_{\boldsymbol{\theta}} \bar{\ell}(\boldsymbol{\theta})] \right) = (1-\rho_t)\mathbf{S}_t + \rho_t \mathbb{E}_{\tilde{\pi}_t^{2-q}}[\nabla^2_{\boldsymbol{\theta}} \bar{\ell}(\boldsymbol{\theta})] \tag{94}$$

$$\mathbf{m}_{t+1} = \mathbf{m}_t - \rho_t \mathbf{S}_{t+1}^{-1} \mathbb{E}_{\pi_t}[\nabla \bar{\ell}(\boldsymbol{\theta})], \tag{95}$$

where $r^2(\boldsymbol{\theta}) = (\boldsymbol{\theta} - \mathbf{m})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta} - \mathbf{m})$ and $\tilde{\pi}^{2-q}$ is the escort distribution of order $2-q$ of $\pi_{\boldsymbol{\lambda}}$. The final form of Eq. 94 follows by canceling the factor $\kappa^{-1}D$ with $\mathbb{E}_{\pi}[r^2(\boldsymbol{\theta})]$, and expressing the expectation with respect to the $(2-q)$-escort.

Heavy tails. q-Gaussians with $q > 1$ resulting from Eq. 81 coincide with Student-t densities with degrees of freedom $\nu = 2/(q-1) - D$ and unbounded support $\mathbb{R}^D$ (further details are reviewed in App. D). We assume $\nu > 2 \iff 1 < q < (D+4)/(D+2)$ so that the covariance exists (and

may optionally require $\nu > 4$ if we need finite kurtosis). We write the Student-t scale matrix as $\mathbf{\Sigma} = \kappa_t |\mathbf{S}|^{-1/\nu} \mathbf{S}^{-1}$ with constant

$$\kappa_t(\nu, D) = \frac{\nu + D}{\nu} \left[(\nu + D)^{D/2} \pi^{D/2} \frac{\Gamma(\nu/2)}{\Gamma((\nu + D)/2)} \right]^{2/\nu},$$

where Γ is the Gamma function. Using the Student-t parametrization $\pi_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \mathcal{T}_\nu(\boldsymbol{\theta} \mid \mathbf{m}, \mathbf{\Sigma})$, we can derive from Eq. 88 that $\mathrm{Var}_{\pi_{\boldsymbol{\lambda}}}(\boldsymbol{\theta}) = \nu/(\nu - 2)\mathbf{\Sigma}$ (see App. D.3).

Thus, from Eq. 89 we obtain

$$\nabla_{\boldsymbol{\mu}} f = \nabla_{\mathbf{m}} f - 2 \frac{\nu - 2}{\nu} \nabla_{\mathbf{\Sigma}} f \mathbf{m}, \tag{96}$$

and, from Eq. 90, since $\partial \mathbf{\Sigma} / \partial \boldsymbol{\mu} = -2/\nu(\nu - 2)\mathbf{m}$ and $\partial \mathbf{\Sigma} / \partial \mathbf{M} = (\nu - 2)/\nu$

$$\nabla_{\mathbf{M}} f = \frac{\nu - 2}{\nu} \nabla_{\mathbf{\Sigma}} f. \tag{97}$$

Again, setting $f = \bar{\ell}$, the chain rule of Eq. 87 results in

$$\begin{aligned} \boldsymbol{\lambda}_{t+1} &= (1 - \rho_t) \boldsymbol{\lambda}_t \\ &\quad - \rho_t \begin{bmatrix} \nabla_{\mathbf{m}} \mathbb{E}_\pi[\bar{\ell}(\boldsymbol{\theta})] - 2 \frac{\nu-2}{\nu} \left(\nabla_{\mathbf{\Sigma}} \mathbb{E}_\pi[\bar{\ell}(\boldsymbol{\theta})] \right) \mathbf{m} \\ \frac{\nu-2}{\nu} \nabla_{\mathbf{\Sigma}} \mathbb{E}_\pi[\bar{\ell}(\boldsymbol{\theta})] \end{bmatrix}. \end{aligned} \tag{98}$$

In App. E we prove the Student-t Bonnet-type identity $\nabla_{\mathbf{m}} \mathbb{E}_\pi[\bar{\ell}(\boldsymbol{\theta})] = \mathbb{E}_\pi[\nabla_{\boldsymbol{\theta}} \bar{\ell}(\boldsymbol{\theta})]$ and the Student-t Price-type identity:

$$\nabla_{\mathbf{\Sigma}} \mathbb{E}_\pi[\bar{\ell}(\boldsymbol{\theta})] = \frac{\nu}{2(\nu + D - 2)} \mathbb{E}_\pi \left[\left(1 + \tfrac{1}{\nu} r^2(\boldsymbol{\theta})\right) \nabla^2_{\boldsymbol{\theta}} \bar{\ell}(\boldsymbol{\theta}) \right]. \tag{99}$$

Similarly to the bounded-support case, we can express the Price-type identity in terms of the $(2 - q)$-escort

$$\nabla_{\mathbf{\Sigma}} \mathbb{E}_\pi[\bar{\ell}(\boldsymbol{\theta})] = \frac{\nu}{2(\nu - 2)} \mathbb{E}_{\tilde{\pi}} \left[\nabla^2_{\boldsymbol{\theta}} \bar{\ell}(\boldsymbol{\theta}) \right], \tag{100}$$

which is a neat fact that, as far as we know, has not been noticed before. Since $q = 1 + 2/(\nu + D)$ and $\pi(\boldsymbol{\theta}) \propto \left(1 + r^2(\boldsymbol{\theta})/\nu\right)^{-(\nu+D)/2}$, with r^2 defined as in the bounded-support case, we can express the $(2 - q)$-escort as a reweighting of the base distribution π:

$$\tilde{\pi}(\boldsymbol{\theta})^{2-q} = \frac{\pi(\boldsymbol{\theta})^{2-q}}{\int \pi(\boldsymbol{\theta})^{2-q} d\boldsymbol{\theta}} \propto \pi(\boldsymbol{\theta}) \left(1 + \tfrac{r^2}{\nu}(\boldsymbol{\theta})\right). \tag{101}$$

For $\nu > 2$,

$$\mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[r^2(\boldsymbol{\theta})] = \frac{D\nu}{\nu - 2} \Rightarrow \mathbb{E}_{\pi_{\boldsymbol{\lambda}}}[1 + \frac{r^2}{\nu}] = \frac{D + \nu - 2}{\nu - 2},$$

and substituting into Eq. 99 yields Eq. 100.

Thus, we obtain the Student-t Newton-like canonical parameter updates that are used in Alg. 2 by solving Eq. 98:

$$\mathbf{S}_{t+1} = (1 - \rho_t)\mathbf{S}_t + \rho_t \mathbb{E}_{\tilde{\pi}_t^{2-q}} \left[\nabla^2_{\boldsymbol{\theta}} \bar{\ell}(\boldsymbol{\theta}) \right], \tag{102}$$

$$\mathbf{m}_{t+1} = \mathbf{m}_t - \rho_t \mathbf{S}_{t+1}^{-1} \mathbb{E}_{\pi_t} \left[\nabla_{\boldsymbol{\theta}} \bar{\ell}(\boldsymbol{\theta}) \right]. \tag{103}$$

Implementation. In both the bounded-support and heavy-tailed cases we exploit escort distributions to mirror the familiar Gaussian stochastic gradient expressions in an elegant way, and to highlight the probabilistic structure that would otherwise be obscured by repeated integration by parts. For efficient sampling in Alg. 1 and Alg. 2, we express $\tilde{\pi}^{2-q}$ as a reweighting of $\pi_{\boldsymbol{\lambda}}$. When $q > 1$ the $(2 - q)$-escort relatively upweights larger values of $r(\boldsymbol{\theta})$ since typical samples dominate $\mathbb{E}_\pi[\nabla \bar{\ell}]$ and the mean update is mode-driven. The scale update in this case is more tail-aware since atypical large-radius samples influence the Hessian estimate. The combined effect is a conservative preconditioning because a larger effective curvature leads to smaller steps in $\mathbf{m}$, but possibly higher variance in the curvature estimator. When $q < 1$ the $(2 - q)$-escort concentrates on the interior of the support. Qualitatively, in this case, the mean is more sensitive to the bounded-support shape, while

the scale update has lower variance, leading to a stable scale estimate, with bounded exploration due to the compact support.

We make additional adjustments to the generalized BLR q-VON updates in Eqs. 102 and 103 and Eqs. 94 and 95, in order to easily implement them as optimizers in modern deep-learning frameworks. Our simplifications are analogous to Khan et al. [2018, App. D]'s and Shen et al. [2024]'s modifications of the BLR that lead to deep-learning-friendly VON and IVON algorithms. First, the loss function is reweighted with an inverse temperature $\zeta > 0$ and scaled by $1/N$, that is, we use $\zeta \frac{1}{N} \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta})$ in Eq. 13. This reweighting allows us to use mini-batch stochastic gradients along with sampling to estimate the expectations (lines 2 of Alg. 1 and Alg. 2). Second, we use a diagonal $\mathbf{S} = \text{diag}(\mathbf{s})$ and reparametrize it via $\mathbf{s} = \zeta(\mathbf{h} + \delta)$ (lines 7), where $\delta > 0$ is the L_2-regularizer, $-\eta_0(\boldsymbol{\theta}) = \frac{\zeta\delta}{2}\|\boldsymbol{\theta}\|^2$ in the generalized Bayesian objective of Eq. 13. The $\mathbf{S}$-updates in Eqs. 94 and 102 are rewritten in terms of $\mathbf{h}$ in lines 5 of the algorithms, and separate learning rates $(1 - \beta_2)$ and α_t are introduced in place of ρ_t. Similar to IVON [Shen et al., 2024], we add a gradient momentum in lines 4. This is not a direct consequence of the theory, but does not change the fixed points of the methods and we found it improves results. The initial $\mathbf{m}$ is set to common initialization schemes, and $\mathbf{S} = \zeta(h_0 + \delta)\mathbf{I}$ where $h_0 > 0$ is a hyperparameter. To stabilize the algorithms, we ensure that the noise added to the parameters remains bounded by clipping the entries of $\mathbf{s}$ to a minimum $s_{\text{min}} > 0$.

F.2 Mean-Field Variants

The constants κ_t, κ and γ used in Alg. 1 and Alg. 2 depend on the dimensionality of the parameter $\boldsymbol{\theta}$. To further adapt the proposed q-VON algorithms to high-dimensional problems, we also propose mean-field variants, where each parameter follows a one-dimensional q-Gaussian. This corresponds to a factorized posterior, that is, $\pi(\boldsymbol{\theta}) = \pi(\theta_1) \cdot \ldots \cdot \pi(\theta_D)$. Mean-field variants of Alg. 1 and Alg. 2 are easily obtained when Ω in the generalized Bayesian variational problem Eq. 13 is replaced with the sum of one-dimensional functions on each parameter.

In practice, the only modification needed to Alg. 1 and Alg. 2 is to set $D = 1$ in the computation of the constants, and compute $\boldsymbol{\sigma}^2$ in lines 8 individually for each s_i rather than taking their product. The sampling step in lines 2 is also modified to sample from a one-dimensional q-Gaussian or Student-t distribution. In the remainder of the paper, we refer to the mean-field q-VON variants as q-VON-MF.

G Experiments

G.1 Bayesian logistic regression

We demonstrate how q-VON-MF with bounded support (Alg. 1, Sec. 4) can be used to approximate the posterior mode in Bayesian logistic regression on the UCI Credit Approval dataset (Fig. 2). Posterior marginals for each feature are computed with MCMC. We compare variational posterior approximations: mean-field Gaussian ($q = 1$) to bounded-support q-Gaussians with $q = -2, -10$.

G.2 Image Classification with Neural Networks

We use q-VON (Alg. 1 and Alg. 2) and the mean-field variant q-VON-MF (described in Sec. 4) to train neural networks on the MNIST, CIFAR-10 and CIFAR-100 datasets. We compare with several baseline learning algorithms on test accuracy, test loss (NLL), expected calibration error Guo et al. [2017], Brier score and the AUROC metric introduced by Osawa et al. [2019]. The baselines are SGD (with momentum) applied to the standard empirical risk minimization objective to find the MAP, variational SGD (VSGD) which is SGD with Gaussian noise injection [Orvieto et al., 2022], and the Improved Variational Online Newton method (IVON) [Shen et al., 2024]. In all experiments, we closely follow the setup of Shen et al. [2024] and use their publicly available code. In particular, we use the same hyperparameters for q-VON and q-VON-MF as IVON. IVON and VSGD target the standard Bayesian objective for Gaussian families with fixed isotropic variance and flexible diagonal variance respectively. Another baseline we consider is q-VSGD, which targets the generalized Bayesian objective but only learns the mean of the approximate posterior distribution. All algorithms have the same complexity and run-time, as each iteration is dominated by the cost of forward and backward propagation through the neural network. The results are in Table 2 for MNIST and Table 1 for CIFAR-10 and CIFAR-100. Below we summarize our main findings.

q-VON consistently improves over q-VSGD and SGD. Both on MNIST and CIFAR-10, we see that q-VON consistently improves over q-VSGD and SGD (MAP). This confirms that learning the shape of the approximate posterior rather than using a fixed shape improves both accuracy and calibration performance, and that the method is superior to pure noise injection approaches [Orvieto et al., 2022].

q-VON can exceed IVON's performance. On the MNIST dataset, the heavy-tailed variant of q-VON ($q > 1$) gives better generalization performance than IVON, whereas the compact support variant ($q < 1$) gives better calibration. On CIFAR-10 and CIFAR-100, q-VON-MF reaches competitive performance with IVON, without tuning any of its hyperparameters.

Posterior sampling helps on the CIFAR datasets. For q-VON and IVON which estimate a distribution over weights, using the posterior predictive distribution generally boosts the performance.

Overall, the experiments demonstrate that the proposed q-VON algorithms and the mean field variants easily scale to large neural networks. The methods significantly improve over MAP estimation and reach competitive performance with the recent state-of-the-art method IVON [Shen et al., 2024]. We used the same setup and hyperparameters from [Shen et al., 2024] (which were tuned for Gaussian posteriors). Using IVON-tuned hyperparameters, q-VON is already competitive; dedicated tuning remains future work.

Tables 1 and 2 report results on deep learning with the proposed q-VON algorithms, as described in Sec. 4, App. F. q-VON ($q < 1$) denotes Alg. 1 with bounded-support q-Gaussian posterior and $q = 0.6$. q-VON ($q > 1$) denotes Alg. 2 with Student-t posterior with $\nu = 5 \cdot 10^5$. q-VON-MF refers to the mean-field variant; each parameter has its own q-Gaussian ($q = 0.6$ for $q < 1$ and $\nu = 5$ for $q > 1$).

Table 1: Results on convolutional neural networks. We evaluate various q-VON variants using the same hyperparameters used in Shen et al. [2024] for IVON. Even without tuning, the results are comparable to IVON, with the heavy-tailed q-VON-MF reaching slightly higher AUROC scores.

Task	Method	Acc. (↑)	NLL (↓)	ECE (↓)	Brier (↓)	AUROC (↑)
CIFAR-10 ResNet-20 270k params	SGD (MAP)	91.2 ± 0.1	0.292 ± 0.001	3.7 ± 0.1	0.133 ± 0.001	91.9 ± 0.1
	VSGD	91.9 ± 0.0	0.310 ± 0.005	4.5 ± 0.1	0.127 ± 0.001	92.1 ± 0.1
	IVON	92.4 ± 0.1	0.259 ± 0.006	3.5 ± 0.2	0.117 ± 0.002	92.4 ± 0.0
	+ sampling	**92.6** ± 0.1	**0.221** ± 0.003	**0.8** ± 0.1	**0.109** ± 0.002	92.6 ± 0.1
	q-VSGD ($q < 1$)	91.9 ± 0.0	0.309 ± 0.008	4.5 ± 0.1	0.128 ± 0.002	92.0 ± 0.1
	q-VON ($q < 1$)	92.3 ± 0.2	0.271 ± 0.008	3.9 ± 0.2	0.119 ± 0.003	92.6 ± 0.2
	+ sampling	92.4 ± 0.2	0.243 ± 0.006	2.6 ± 0.2	0.113 ± 0.003	**92.7** ± 0.2
	q-VON-MF ($q > 1$)	92.4 ± 0.0	0.270 ± 0.003	3.8 ± 0.1	0.118 ± 0.001	92.5 ± 0.2
	+sampling	92.5 ± 0.0	0.229 ± 0.002	1.5 ± 0.1	0.110 ± 0.001	**92.7** ± 0.1
CIFAR-100 PreResNet-110 4M params	SGD (MAP)	73.0 ± 0.2	1.270 ± 0.009	14.1 ± 0.1	0.410 ± 0.002	86.1 ± 0.1
	VSGD	73.8 ± 0.2	1.319 ± 0.009	15.5 ± 0.2	0.410 ± 0.003	86.1 ± 0.1
	IVON	75.1 ± 0.3	1.150 ± 0.012	13.8 ± 0.3	0.382 ± 0.004	86.8 ± 0.2
	+ sampling	**75.7** ± 0.2	**0.880** ± 0.006	**2.9** ± 0.2	**0.336** ± 0.003	87.0 ± 0.2
	q-VSGD ($q < 1$)	73.3 ± 0.1	1.326 ± 0.001	15.6 ± 0.1	0.416 ± 0.002	86.1 ± 0.2
	q-VON ($q < 1$)	74.6 ± 0.2	1.216 ± 0.008	14.0 ± 0.1	0.391 ± 0.002	86.3 ± 0.1
	+ sampling	74.7 ± 0.2	1.189 ± 0.008	13.3 ± 0.1	0.387 ± 0.002	86.4 ± 0.1
	q-VON-MF ($q > 1$)	74.9 ± 0.1	1.204 ± 0.008	14.4 ± 0.1	0.389 ± 0.002	86.7 ± 0.1
	+ sampling	75.4 ± 0.1	0.930 ± 0.005	5.6 ± 0.1	0.344 ± 0.001	**87.1** ± 0.1

Table 2: Additional results on MNIST. Heavy-tailed posteriors can improve generalization performance, whereas truncated ones can improve calibration.

Task	Method	Acc. (↑)	Test Loss (↓)	ECE (↓)	Brier (↓)	AUROC (↑)
MNIST MLP 550k params	SGD (MAP)	98.69 ± 0.01	0.043 ± 0.000	0.37 ± 0.02	0.020 ± 0.000	97.87 ± 0.07
	VSGD	98.75 ± 0.03	0.048 ± 0.000	0.60 ± 0.02	0.020 ± 0.000	97.68 ± 0.05
	IVON	98.74 ± 0.02	0.048 ± 0.001	0.60 ± 0.02	0.020 ± 0.000	97.52 ± 0.10
	+ sampling	98.72 ± 0.01	0.050 ± 0.000	1.53 ± 0.02	0.022 ± 0.000	97.88 ± 0.05
	q-VSGD ($q < 1$)	98.62 ± 0.02	0.045 ± 0.001	0.46 ± 0.01	0.021 ± 0.000	97.94 ± 0.06
	q-VON ($q < 1$)	98.69 ± 0.02	**0.042** ± 0.000	**0.40** ± 0.03	0.020 ± 0.000	98.04 ± 0.05
	+ sampling	98.69 ± 0.02	**0.042** ± 0.000	**0.40** ± 0.04	0.020 ± 0.000	98.04 ± 0.05
	q-VSGD ($q > 1$)	98.59 ± 0.01	0.045 ± 0.000	0.48 ± 0.02	0.022 ± 0.000	97.97 ± 0.05
	q-VON ($q > 1$)	**98.81** ± 0.03	0.043 ± 0.000	0.60 ± 0.01	**0.019** ± 0.000	97.84 ± 0.11
	+ sampling	98.76 ± 0.02	**0.042** ± 0.000	0.85 ± 0.01	0.020 ± 0.000	**98.05** ± 0.09
	q-VON-MF ($q < 1$)	98.66 ± 0.03	0.052 ± 0.001	0.60 ± 0.02	0.021 ± 0.000	97.19 ± 0.16
	+ sampling	98.62 ± 0.02	0.053 ± 0.001	1.60 ± 0.02	0.024 ± 0.000	97.82 ± 0.04
	q-VON-MF ($q > 1$)	98.76 ± 0.02	0.049 ± 0.001	0.59 ± 0.02	0.020 ± 0.000	97.14 ± 0.07
	+sampling	98.67 ± 0.01	0.054 ± 0.001	1.90 ± 0.02	0.023 ± 0.000	97.88 ± 0.05

G.3 Empirical Verification of Convergence

A convergence analysis of BLR/VON-type methods in the generality considered here is still largely open. Existing results cover narrower settings: Wu and Gardner [2024] prove non-asymptotic rates for stochastic natural gradient (NG) VI in conjugate exponential family models and show that for non-conjugate log-concave likelihoods, the negative ELBO need not be convex in the expectation-parameter coordinates used by canonical NGVI. Guilmeau et al. [2025] instead study Bregman-proximal algorithms for regularized Rényi-divergence minimization over exponential families, with convergence guarantees based on relative smoothness with respect to the log-partition geometry. These assumptions are different from ours: our geometry is induced by a density-space Fenchel-Young discrepancy, which defines the generalized maxent family and the corresponding BLR update. A convergence theorem for the Fenchel-Young BLR would be a substantial project in its own right and is outside the scope of this paper. Instead, we provide empirical evidence that the implemented generalized updates move toward lower values of their corresponding variational objectives.

We use 2D Bayesian logistic regression as a minimal nontrivial diagnostic, with bounded-support, Gaussian, and Student-t variants. The plots in Fig. 3 show slices of the optimizer-specific generalized objective over one mean-dispersion coordinate, with the remaining coordinates fixed at its final value. Each panel recomputes the objective associated with that posterior family and parametrization, because the Fenchel-Young discrepancy, sampling law, and scale-precision map change with q and with the full versus mean-field construction. All trajectories move toward low-objective regions and stabilize near the minimum.

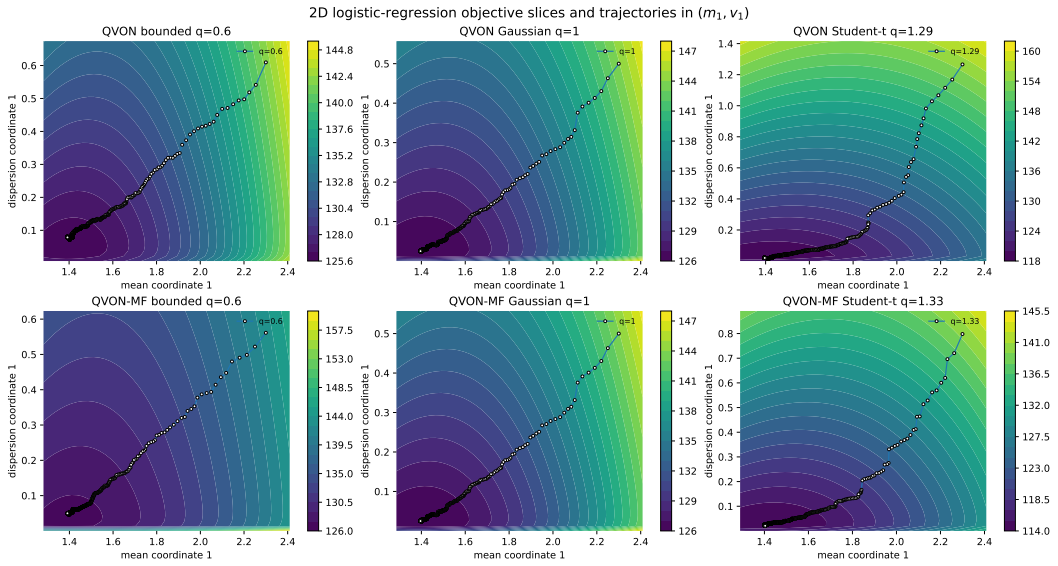


Figure 3: Optimizer-specific convergence on 2D Bayesian logistic regression. Each panel shows the objective associated with an optimizer and posterior family. The plots show slices of the optimizer-specific generalized objective over one mean-dispersion coordinate, with the remaining coordinate fixed at its final value. The trajectories consistently approach low-objective regions for bounded-support, Gaussian, and Student-t variants, for both full and mean-field implementations. The same $\nu = 5$ in both Student-t panels corresponds to different q's because of the dependence on D.

H Varying q

Different choices of q change the variational objective, the scale-precision map, and the stochastic Hessian estimator, so iteration counts across q's are not directly comparable. Nevertheless, q-VON retains the fast VON-style optimizer structure: a Newton-like mean update preconditioned by a learned diagonal precision, implemented with minibatch stochastic gradients, momentum, and Hessian-diagonal estimates rather than explicit second-order computations [Khan et al., 2018, Osawa et al., 2019, Shen et al., 2024]. Since q-VON changes only the sampling law, the precision constants, and the corresponding diagonal Stein estimator, we expect similar fast behavior.

H.1 1D Logistic Regression

In Fig. 4 we use a synthetic dataset of size $N = 256$, with covariates $x_i \sim \mathcal{N}(0, 1)$, labels $y_i \sim \text{Bernoulli}(\sigma(2x_i))$, and Gaussian prior $w \sim \mathcal{N}(0, 0.1)$. For each q, q-VON is run for 3000 steps and repeated over 5 seeds. Points show the average posterior mean or variance, and error bars show 95% confidence intervals. Horizontal reference lines show the exact Bayesian posterior quantities for the ordinary Gaussian-prior model. Exact references are approximated by evaluating the negative log-posterior $\ell(w) + w^2/0.2$ on a grid.

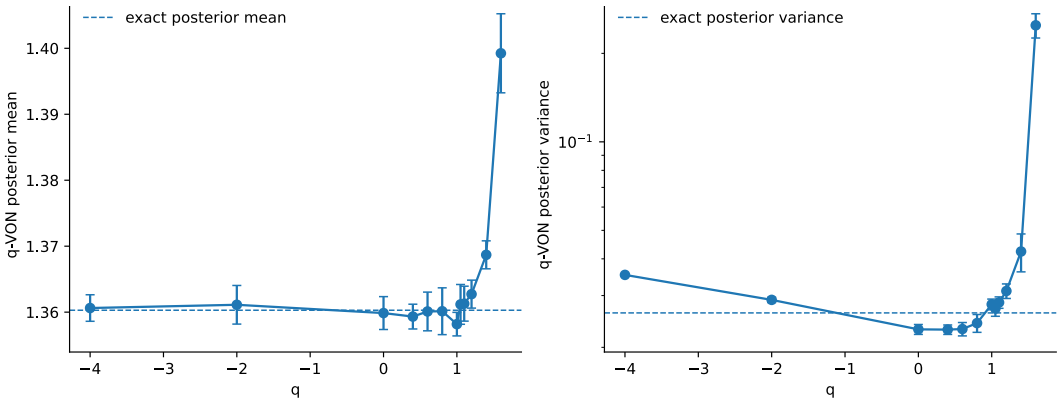


Figure 4: 1D Bayesian logistic regression q-sweep with q-VON.

H.2 MNIST

In Table 3 we report the results of varying q on MNIST training with q-VON-MF, again using the hyperparameters obtained from previous tuning with IVON.

Table 3: MNIST q-sweep for q-VON-MF. Results are mean ± s.e. over 5 seeds.

Task	Method	Acc. (↑)	Test Loss (↓)	ECE (↓)	Brier (↓)	AUROC (↑)
MNIST MLP 550k params	q-VON-MF ($q = -10$)	81.42 ± 15.77	0.514 ± 0.412	1.35 ± 0.59	0.208 ± 0.165	88.55 ± 6.79
	q-VON-MF ($q = -8$)	97.67 ± 0.03	0.086 ± 0.001	0.88 ± 0.01	0.036 ± 0.000	96.26 ± 0.18
	q-VON-MF ($q = -6$)	98.15 ± 0.08	0.070 ± 0.002	0.86 ± 0.04	0.029 ± 0.001	96.79 ± 0.12
	q-VON-MF ($q = -4$)	98.66 ± 0.03	0.053 ± 0.001	0.72 ± 0.01	0.022 ± 0.000	97.67 ± 0.12
	q-VON-MF ($q = -2$)	**98.75** ± 0.01	0.048 ± 0.001	0.61 ± 0.03	**0.020** ± 0.000	97.52 ± 0.05
	q-VON-MF ($q = -1$)	98.66 ± 0.02	0.047 ± 0.000	0.60 ± 0.02	0.021 ± 0.000	97.93 ± 0.07
	q-VON-MF ($q = 0$)	98.60 ± 0.01	0.046 ± 0.001	0.54 ± 0.02	0.021 ± 0.000	97.98 ± 0.05
	q-VON-MF ($q = 0.4$)	98.59 ± 0.03	0.047 ± 0.001	0.51 ± 0.02	0.022 ± 0.000	97.88 ± 0.05
	q-VON-MF ($q = 0.6$)	98.56 ± 0.02	0.047 ± 0.001	0.50 ± 0.02	0.022 ± 0.000	98.01 ± 0.03
	q-VON-MF ($q = 0.8$)	98.57 ± 0.02	0.046 ± 0.000	0.48 ± 0.02	0.022 ± 0.000	97.91 ± 0.05
	q-VON-MF ($q = 0.95$)	98.55 ± 0.04	0.047 ± 0.001	0.48 ± 0.03	0.022 ± 0.000	97.98 ± 0.06
	q-VON-MF ($q = 1$)	98.57 ± 0.01	0.046 ± 0.001	**0.45** ± 0.02	0.022 ± 0.000	98.02 ± 0.04
	q-VON-MF ($q = 1.000004$)	98.53 ± 0.02	0.046 ± 0.000	0.50 ± 0.02	0.022 ± 0.000	98.03 ± 0.04
	q-VON-MF ($q = 1.02$)	98.55 ± 0.03	0.046 ± 0.001	0.48 ± 0.01	0.022 ± 0.000	97.95 ± 0.08
	q-VON-MF ($q = 1.05$)	98.53 ± 0.01	0.046 ± 0.000	0.49 ± 0.01	0.022 ± 0.000	98.05 ± 0.07
	q-VON-MF ($q = 1.1$)	98.56 ± 0.01	0.046 ± 0.000	0.50 ± 0.02	0.022 ± 0.000	98.03 ± 0.03
	q-VON-MF ($q = 1.2$)	98.53 ± 0.02	0.046 ± 0.000	0.48 ± 0.03	0.022 ± 0.000	98.03 ± 0.03
	q-VON-MF ($q = 4/3$)	98.51 ± 0.01	0.046 ± 0.000	0.47 ± 0.02	0.022 ± 0.000	98.05 ± 0.04
	q-VON-MF ($q = 1.5$)	98.54 ± 0.01	0.046 ± 0.001	0.48 ± 0.03	0.022 ± 0.000	98.00 ± 0.03
	q-VON-MF ($q = 1.57143$)	98.55 ± 0.02	0.046 ± 0.001	0.48 ± 0.01	0.022 ± 0.000	97.99 ± 0.04
	q-VON-MF ($q = 1.6$)	98.55 ± 0.03	0.046 ± 0.000	0.48 ± 0.03	0.022 ± 0.000	98.02 ± 0.07
	q-VON-MF ($q = 1.625$)	98.54 ± 0.04	0.047 ± 0.001	0.48 ± 0.04	0.022 ± 0.000	98.01 ± 0.06
	q-VON-MF ($q = 1.64516$)	98.55 ± 0.03	**0.046** ± 0.000	0.48 ± 0.04	0.022 ± 0.000	**98.07** ± 0.06
	q-VON-MF ($q = 1.65574$)	98.55 ± 0.02	0.046 ± 0.001	0.46 ± 0.03	0.022 ± 0.000	97.93 ± 0.05
	q-VON-MF ($q = 1.66445$)	98.57 ± 0.03	0.046 ± 0.000	0.47 ± 0.03	0.022 ± 0.000	97.99 ± 0.07

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's scope: a Fenchel-Young or mixed-Bregman generalization of the BLR, q-Gaussian examples, and experiments illustrating controllable support and tail behavior.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly limits connections with robust Bayes, state-of-the-art, scale, and convergence claims, and mentions several directions as future work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The main theoretical claims state their assumptions, and the supporting proofs and derivations are given in the appendices.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper gives the algorithms, datasets, baselines, metrics, and experimental setup needed to reproduce the reported results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The datasets and baseline code are public and cited, but a complete anonymized code release is not included with this submission. We will release the code upon acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The main text summarizes the experiments, and the appendix gives the datasets, baselines, metrics, and hyperparameter choices.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The tables report variability across runs, and the 1D q-sweep reports confidence intervals across optimizer seeds.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper discusses algorithmic complexity and per-iteration cost, but does not provide detailed hardware, memory, or wall-clock compute. Scale is not one of the paper's main considerations. 1 GPU is sufficient to reproduce our experiments in a few days.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our work uses standard public benchmarks and theoretical analysis, and we are not aware of any deviation from the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is foundational work on variational optimization and does not introduce high-risk models, or application-specific datasets.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk models, scraped datasets, or other assets requiring special safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The paper cites the public datasets and baseline code used in the experiments, but does not separately list their licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce a new dataset, benchmark, or pretrained model.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject experiments, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as a method, model, evaluator, or original component of the research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.