The Bayesian Learning Rule Beyond KL Geometry

Sophia Sklaviadis
with Wu Lin, Andre F. T. Martins, Mario A. T. Figueiredo,
Thomas Moellenhoff, M. Emtiyaz Khan

November 12, 2026

Bayesian Updating Beyond KL Geometry

prior potential	$+$	data contribution	$\longrightarrow$	posterior
η_0	$+$	$(-\ell_{\mathcal{D}})$	$\longrightarrow$	$\pi^\star(\boldsymbol{\theta}\|\mathcal{D}) \in \partial\Omega^*(\eta_0 - \ell_{\mathcal{D}})$
$\log \pi_0$	$+$	$\log p(\mathcal{D} \mid \boldsymbol{\theta})$	$\longrightarrow$	$\dfrac{\exp\left(\log \pi_0 + \log p(\mathcal{D} \mid \boldsymbol{\theta})\right)}{\int_\Theta \exp\left(\log \pi_0 + \log p(\mathcal{D} \mid \boldsymbol{\theta})\right) dh_\Theta}$ } $\partial\Omega^*_{\mathrm{Sh}}$ Shannon dual map

Bayes' rule: additive score-space update.

Fenchel-Young Bayes: replace the normalized exponential map by $\partial\Omega^*$.

Posteriors as unconstrained optimizers of the FY free energy:

$$\pi^\star \in \arg\min_{\pi\in\mathcal{P}} \mathbb{E}_\pi[\ell_{\mathcal{D}}(\boldsymbol{\theta})] + \mathbb{D}_\Omega(\eta_0, \pi)$$

$$\mathbb{D}_\Omega(\eta_0, \pi) = \Omega(\pi) + \Omega^*(\eta_0) - \mathbb{E}_\pi[\eta_0]$$

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

$$= \arg\min_{\pi\in\mathcal{P}} \mathbb{D}_\Omega\left(\eta_0 - \ell(\mathbf{y};\boldsymbol{\theta}); \pi\right) + \Omega^*(\eta_0) - \Omega^*\left(\eta_0 - \ell(\mathbf{y};\boldsymbol{\theta})\right)$$

The Fenchel-Young Bayesian Learning Rule

Maxent families: $\mathbf{T} : \Theta \to \mathbb{R}^P$, $\boldsymbol{\lambda} \in \mathbb{R}^P$

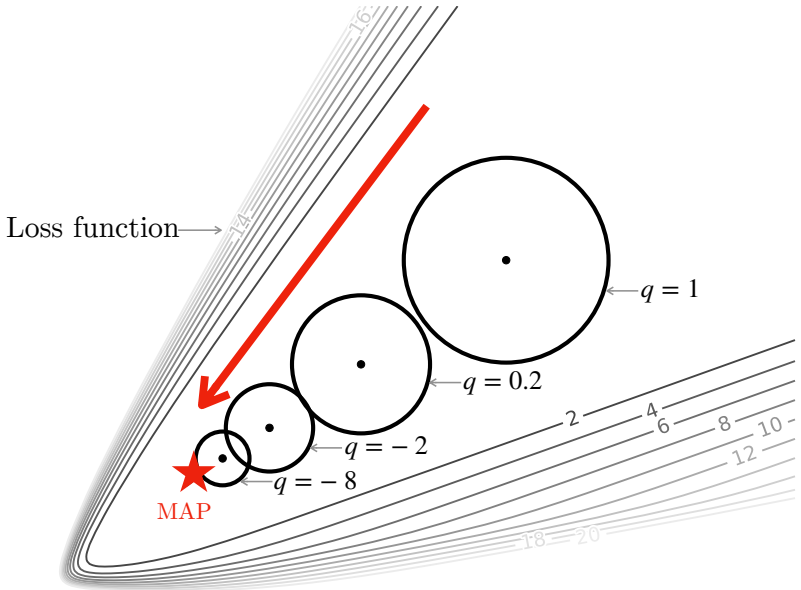
$$\eta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \langle \boldsymbol{\lambda}, \mathbf{T}(\boldsymbol{\theta}) \rangle$$
$$\mathcal{Q}_{\Omega,\mathbf{T}} = \{\pi_{\boldsymbol{\lambda}} : \pi_{\boldsymbol{\lambda}} \in \partial\Omega^*(\eta_{\boldsymbol{\lambda}}),\ \boldsymbol{\lambda} \in \Lambda\}$$

Dual coordinates:

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

Generalized Fisher metric:

$$\mathbf{F}(\boldsymbol{\lambda}) := \nabla^2_{\boldsymbol{\lambda}}\Omega^*(\eta_{\boldsymbol{\lambda}})$$



Fenchel-Young BLR:

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

When $\Omega = -\mathcal{H}$ we recover the BLR.

Conjugacy as closure under translation:

$$\exists\, \boldsymbol{\lambda}' \in \Lambda :\ \eta_{\boldsymbol{\lambda}_0} - \ell_{\mathcal{D}} = \eta_{\boldsymbol{\lambda}'} + \text{const.} \iff \ell_{\mathcal{D}}(\boldsymbol{\theta}) = \langle \boldsymbol{\lambda}_0 - \boldsymbol{\lambda}', \mathbf{T}(\boldsymbol{\theta}) \rangle + \text{const.}$$

* Linearize the expected loss in $\mathbb{E}_{\pi}[\mathbf{T}]$, then apply the conjugate update.

Example: Tsallis Negentropy and q-VON

Quadratic $\mathbf{T}$ with Tsallis $\Omega \to$ scalable optimizers with bounded-support or heavy-tailed weight distributions.

Tsallis negentropy:

$$\Omega_{2-q}(\pi) = \begin{cases} \dfrac{\mathbb{E}_\pi[\pi(\boldsymbol{\theta})^{1-q}] - 1}{(2-q)(1-q)}, & q \neq 1 \\ \mathbb{E}_\pi[\log \pi(\boldsymbol{\theta})], & q = 1 \end{cases}$$

Quadratic statistics: $\boldsymbol{\theta} \in \mathbb{R}^D$

$$\mathbf{T}(\boldsymbol{\theta})^\top = (\boldsymbol{\theta}, \boldsymbol{\theta}\boldsymbol{\theta}^\top)$$
$$\boldsymbol{\lambda}^\top = (\mathbf{S}\mathbf{m}, -\tfrac{1}{2}\mathbf{S})$$

This Ω yields a family of q-Gaussians that recovers Gaussians at $q = 1$:

$$q < 1: \text{ bounded support,}$$
$$q > 1: \text{ Student-}t, \quad \nu = 2/(q-1) - D > 2$$

Newton-like canonical updates:

$$\mathbf{S}_{t+1} = (1-\rho_t)\mathbf{S}_t + \rho_t \mathbb{E}_{\tilde{\pi}^{2-q}_{\boldsymbol{\lambda}_t}}\left[\nabla^2_{\boldsymbol{\theta}}\left(\ell(\boldsymbol{\theta}) - \eta_{\boldsymbol{\lambda}_0}(\boldsymbol{\theta})\right)\right], \; \tilde{\pi}^{2-q}_{\boldsymbol{\lambda}_t}(\boldsymbol{\theta}) = \frac{\pi_{\boldsymbol{\lambda}_t}(\boldsymbol{\theta})^{2-q}}{\int_{\mathbb{R}^D} \pi_{\boldsymbol{\lambda}_t}(\boldsymbol{\theta}')^{2-q} d\boldsymbol{\theta}'}$$

$$\mathbf{m}_{t+1} = \mathbf{m}_t - \rho_t \mathbf{S}^{-1}_{t+1} \mathbb{E}_{\pi_{\boldsymbol{\lambda}_t}}\left[\nabla_{\boldsymbol{\theta}}\left(\ell(\boldsymbol{\theta}) - \eta_{\boldsymbol{\lambda}_0}(\boldsymbol{\theta})\right)\right]$$