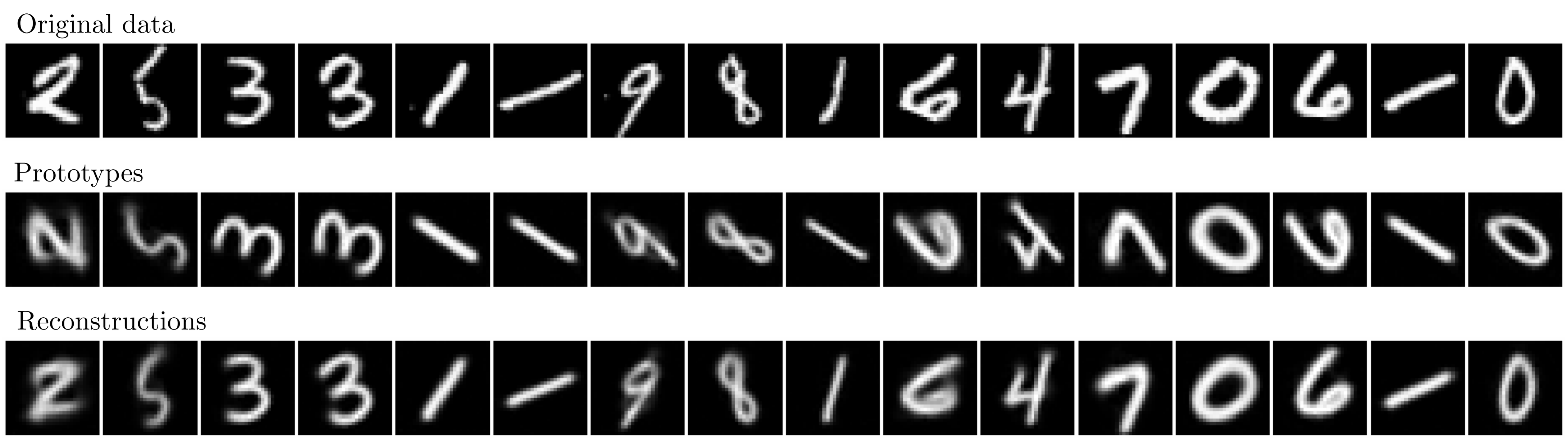
We perform exact inference and hyperparameter optimisation in Bayesian linear models with millions of parameters

Take a picture to see the full paper.

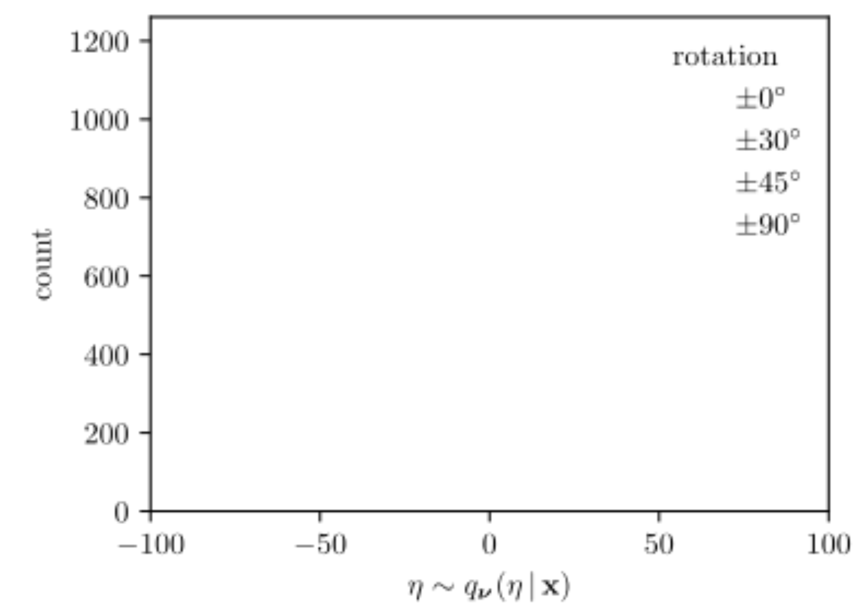
Sampling-based inference for large linear models, with application to linearised Laplace

Javier Antorán*, Shreyas Padhy,* Riccardo Barbano, Eric Nalisnick, David Janz, and José Miguel Hernández-Lobato

We learn representations of MNIST digits that are invariant to rotations and successfully reconstruct the original digits:

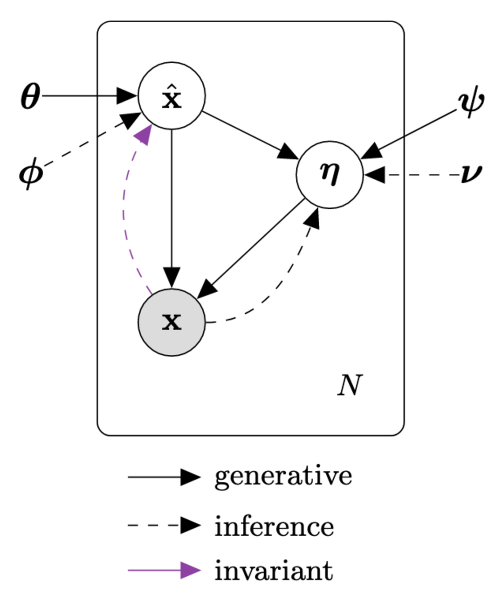


As digits are rotated more, we learn to predict larger angles:



Our model:

$$\hat{\mathbf{x}} \sim p_{\boldsymbol{\theta}}(\hat{\mathbf{x}})\,, \tag{1}$$
$$\boldsymbol{\eta} \sim p_{\boldsymbol{\psi}}(\boldsymbol{\eta}\,|\,\hat{\mathbf{x}})\,, \tag{2}$$
$$\mathbf{x} \sim p(\mathbf{x}\,|\,\boldsymbol{\eta},\,\hat{\mathbf{x}})\,, \tag{3}$$



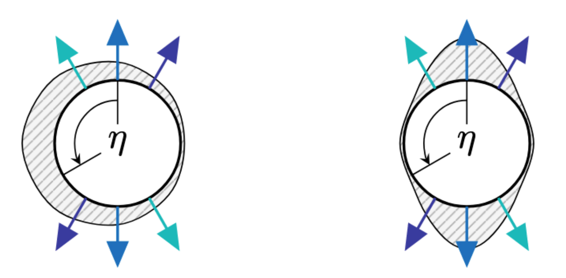
Affine transformation parameterization:

$$\mathcal{T}_{\boldsymbol{\eta}}(\hat{\mathbf{x}}) = T_{\boldsymbol{\eta}} \cdot \hat{\mathbf{x}}, \quad T_{\boldsymbol{\eta}} = \exp\left(\sum_i \eta_i G_i\right) \tag{4}$$

Our training objective:

$$\log p(\mathbf{x}) = \log \iint p(\mathbf{x},\,\hat{\mathbf{x}},\,\boldsymbol{\eta})\,d\boldsymbol{\eta}\,d\hat{\mathbf{x}} \tag{5}$$
$$= \log \mathop{\mathbb{E}}_{q_{\boldsymbol{\nu}}(\boldsymbol{\eta}\,|\,\mathbf{x})\,q_{\boldsymbol{\phi}}(\hat{\mathbf{x}}\,|\,\mathbf{x})}\left[\frac{p(\mathbf{x}\,|\,\hat{\mathbf{x}},\,\boldsymbol{\eta})\;p_{\boldsymbol{\psi}}(\boldsymbol{\eta}\,|\,\hat{\mathbf{x}})\;p_{\boldsymbol{\theta}}(\hat{\mathbf{x}})}{q_{\boldsymbol{\nu}}(\boldsymbol{\eta}\,|\,\mathbf{x})\;q_{\boldsymbol{\phi}}(\hat{\mathbf{x}}\,|\,\mathbf{x})}\right] \tag{6}$$
$$\geq \mathop{\mathbb{E}}_{q_{\boldsymbol{\nu}}\,q_{\boldsymbol{\phi}}}\left[\log p(\mathbf{x}\,|\,\hat{\mathbf{x}},\,\boldsymbol{\eta})\right] - \mathop{\mathbb{E}}_{q_{\boldsymbol{\phi}}}\left[D_{\mathrm{KL}}\left(q_{\boldsymbol{\nu}}\,||\,p_{\boldsymbol{\psi}}\right)\right] - D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}\,||\,p_{\boldsymbol{\theta}}\right) \equiv -\mathcal{L}\left(\boldsymbol{\theta},\,\boldsymbol{\psi},\,\boldsymbol{\phi},\,\boldsymbol{\nu}\right) \tag{7}$$

Conjecture 1: $p_{\boldsymbol{\psi}}(\eta\,|\,\hat{\mathbf{x}})$ must be sufficiently flexible.

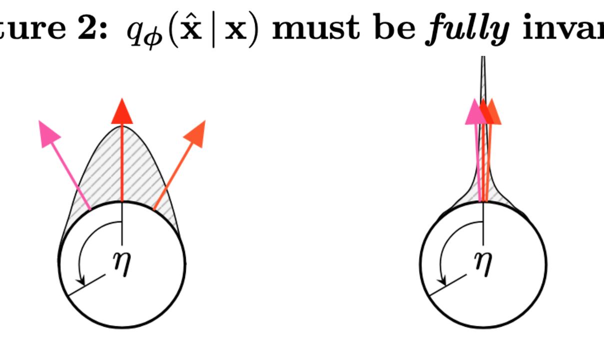


(a) Simple $p_{\boldsymbol{\psi}}(\eta\,|\,\hat{\mathbf{x}})$ (b) Flexible $p_{\boldsymbol{\psi}}(\eta\,|\,\hat{\mathbf{x}})$

$p_{\boldsymbol{\psi}}(\eta\,|\,\hat{\mathbf{x}})$—and $p(\eta\,|\,\hat{\mathbf{x}})$—.

$\mathbf{x}$	$\hat{\mathbf{x}}$	$p(\eta\,|\,\mathbf{x},\,\hat{\mathbf{x}})$
8	8	$0.5 \cdot \delta(\eta - 0°) + 0.5 \cdot \delta(\eta - 180°)$
8	8	$0.5 \cdot \delta(\eta - 30°) + 0.5 \cdot \delta(\eta + 150°)$
8	8	$0.5 \cdot \delta(\eta + 30°) + 0.5 \cdot \delta(\eta - 150°)$

Conjecture 2: $q_{\boldsymbol{\phi}}(\hat{\mathbf{x}}\,|\,\mathbf{x})$ must be *fully* invariant w.r.t η.

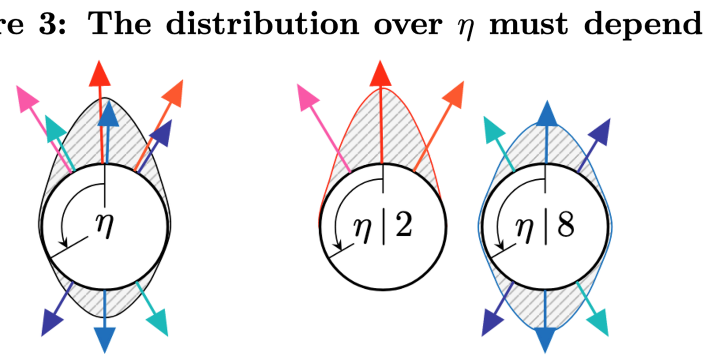


(a) $\rho \geq 60°$ (b) $60° > \rho$

$p_{\boldsymbol{\psi}}(\eta\,|\,\hat{\mathbf{x}})$—and $p(\eta\,|\,\hat{\mathbf{x}})$—.

	$\rho \geq 60°$		$60° > \rho$	
$\mathbf{x}$	$\hat{\mathbf{x}}$	$p(\eta\,|\,\mathbf{x},\,\hat{\mathbf{x}})$	$\hat{\mathbf{x}}$	$p(\eta\,|\,\mathbf{x},\,\hat{\mathbf{x}})$
2	2	$\delta(\eta - 0°)$	2	$\delta(\eta - 0°)$
2	2	$\delta(\eta - 30°)$	2	$\delta(\eta - 0°)$
2	2	$\delta(\eta + 30°)$	2	$\delta(\eta - 0°)$

Conjecture 3: The distribution over η must depend on $\hat{\mathbf{x}}$.



(a) η (b) $\eta\,|\,\hat{\mathbf{x}}$

$p_{\boldsymbol{\psi}}(\eta)$ or $p_{\boldsymbol{\psi}}(\eta\,|\,\hat{\mathbf{x}})$—and $p(\eta)$ or $p(\eta\,|\,\hat{\mathbf{x}})$—.

$\mathbf{x}$	$\hat{\mathbf{x}}$	$p(\eta\,|\,\mathbf{x},\,\hat{\mathbf{x}})$
2	2	$\delta(\eta - 0°)$
2	2	$\delta(\eta - 30°)$
2	2	$\delta(\eta + 30°)$
8	8	$0.5 \cdot \delta(\eta - 0°) + 0.5 \cdot \delta(\eta - 180°)$
8	8	$0.5 \cdot \delta(\eta - 30°) + 0.5 \cdot \delta(\eta + 150°)$
8	8	$0.5 \cdot \delta(\eta + 30°) + 0.5 \cdot \delta(\eta - 150°)$