

Gradient Estimation for Mixture Variational Inference

We study and unify existing estimators, improve and extend them, and develop new ones.

Javier Burroni & Daniel Sheldon

University of Massachusetts Amherst Journal of Machine Learning Research 27 (2026) javier.burroni@gmail.com



Read the paper

1. Setup

Black-box VI with a **mixture** as the variational family (learned weights):

$$q_{\theta}(z) = \sum_{k=1}^K \pi_{k,\theta} q_{k,\theta}(z)$$

Goal: low-variance, cheap estimates of $\nabla_{\theta} \mathbb{E}_{q_{\theta}}[\mathcal{L}_{\theta}(z)]$.

The obstacle: sampling draws a component $k \sim \text{Cat}(\pi_{1:K,\theta})$ first. This discrete step blocks the reparameterization trick. Score-function gradients are noisy; stratifying over components costs K model evaluations, and $K_1 \times \dots \times K_T$ for chains of mixtures.

3. The estimators at a glance

Estimator	Gradient estimate	Evals of $\mathcal{L}$	Time overhead*	Requires
● Stratified	$\sum_k \nabla_{\theta} (\pi_{k,\theta} \mathcal{L}_{\theta}(T_{k,\theta}(\epsilon_k)))$, $\epsilon_{1:K} \stackrel{iid}{\sim} q_{\text{base}}$	K per sample	+13–16%	Reparam. components
● Post-stratified	$\sum_k q_{\theta}(k z) [\nabla_{\theta} \mathcal{L}_{\theta}(T_{k,\theta}(\epsilon)) _{\epsilon=T_{k,\theta}^{-1}(z)} + \mathcal{L}_{\theta}(z) \nabla_{\theta} \ln \pi_{k,\theta}]$	1	+6%	Reparam. components
● J&K	$\sum_k q_{\theta}(k z) \nabla_{\theta} \mathcal{L}_{\theta}(T_{k,\theta}(\epsilon)) _{\epsilon=T_{k,\theta}^{-1}(z)} + V_{\theta}(z) \cdot \nabla_z \mathcal{L}_{\theta}(z)$	1	+6%	Diag. Gaussians, softmax weights
● Implicit reparam.	$-\nabla_{\theta} S_{\theta}(z) (\nabla_z S_{\theta}(z))^{-1} \nabla_z \mathcal{L}_{\theta}(z) + \nabla_{\theta} \mathcal{L}_{\theta}(z)$	1 + D passes of S	+15–65%	Standardization map S_{θ}
● Score function	$\mathcal{L}_{\theta}(z) \sum_k q_{\theta}(k z) \nabla_{\theta} \ln (\pi_{k,\theta} q_{k,\theta}(z)) + \nabla_{\theta} \mathcal{L}_{\theta}(z)$	1	baseline	Differentiable densities

Responsibility $q_{\theta}(k | z) = \pi_{k,\theta} q_{k,\theta}(z) / q_{\theta}(z)$; $T_{k,\theta}$ reparameterizes component k . All are unbiased; the paper derives each in one common notation. *Time per epoch vs. score function, VAEs with $K = 5$, 1 particle (diagonal Gaussian $\rightarrow$ dense Student’s t).

4. Post-stratified estimator NEW DERIVATION

Draw one $z \sim q_{\theta}$. Take the gradient we *would* have had if z came from each component; average by responsibility. No transport equations needed.

One evaluation of $\mathcal{L}$. With $\tilde{\epsilon}_k = T_{k,\theta}^{-1}(z)$, autodiff on the surrogate

$$S_{\theta}^{\text{PS}}(\tilde{\epsilon}_{1:K}) = \mathcal{L}_{\theta}(\sum_k \tilde{q}_{\theta}(k | z) T_{k,\theta}(\tilde{\epsilon}_k))$$

gives the reparam. term; add score term $\mathcal{L}_{\theta}(z) \sum_k q_{\theta}(k | z) \nabla_{\theta} \ln \pi_{k,\theta}$.

Proposition: variance reduction

$$\hat{g}^{\text{PS}}(z) = \mathbb{E}[\hat{g}^{R\kappa} | z] \implies \text{Var}[\hat{g}^{\text{PS}}] \leq \text{Var}[\hat{g}^{R\kappa}]$$

It is the conditional expectation of the randomized-components estimator $\hat{g}^{R\kappa}$ (sample k , then z).

5. Implicit reparameterization EXTENDED · FASTER

Given a standardization map $S_{\theta}(z) \sim q_{\text{base}}$, differentiate it implicitly:

$$\hat{g}^{\text{IR}}(z) = -\nabla_{\theta} S_{\theta}(z) (\nabla_z S_{\theta}(z))^{-1} \nabla_z \mathcal{L}_{\theta}(z) + \nabla_{\theta} \mathcal{L}_{\theta}(z)$$

Broader. S is built from conditional CDFs; we derive them for mixtures of multivariate **Student’s t** and **Cauchy** components.

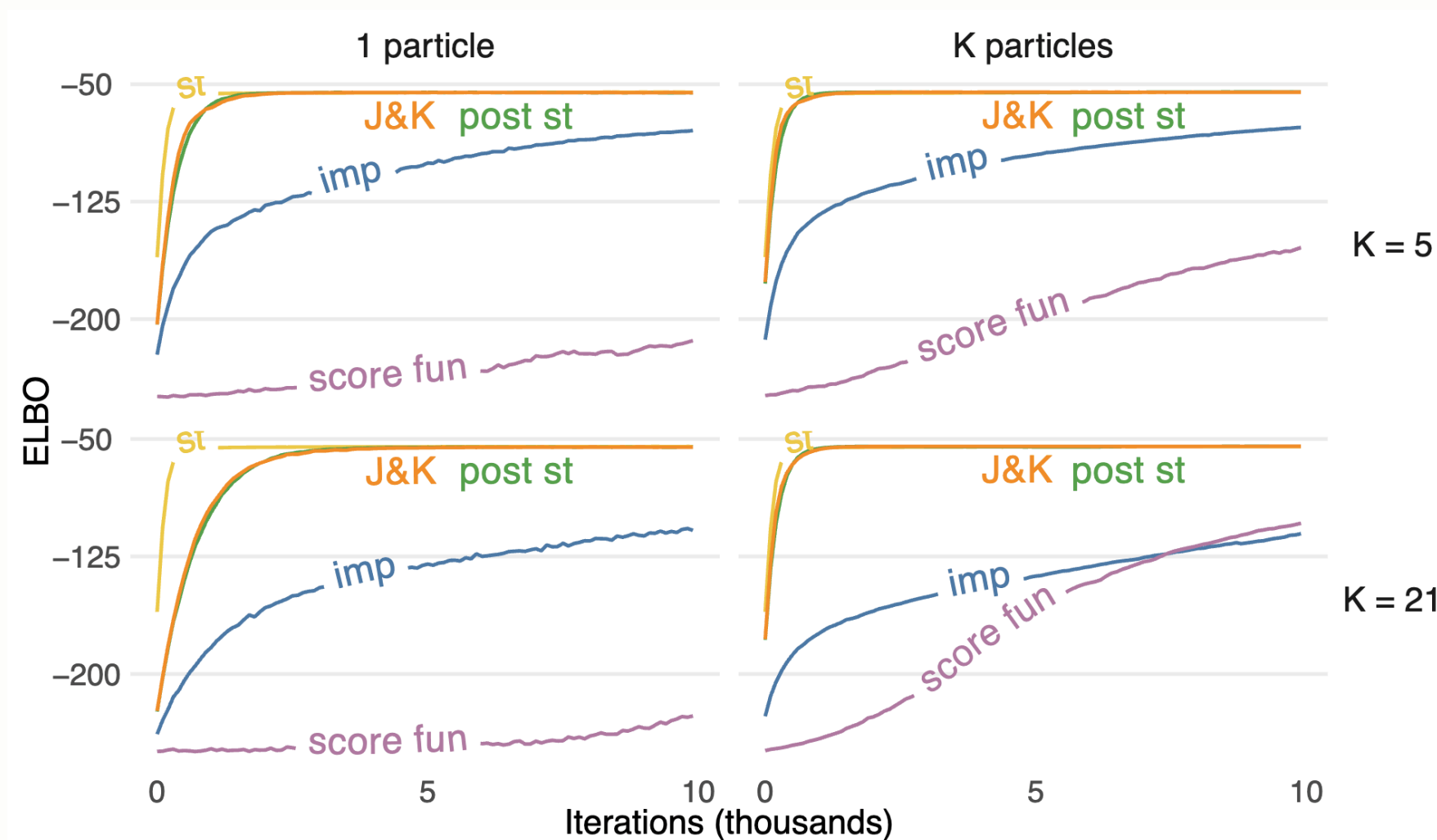
Cheaper. Cholesky factors $\Sigma_k = L_k L_k^{\top}$ cut dense-covariance cost from $O(D^5) \rightarrow O(D^3)$, yet it remains the slowest single-sample method.

Algorithm 1 Compute S for a mixture of dense Gaussians

1	Component weights $W \leftarrow \text{Weights}(z, \pi)$	$O(KD^2)$
2	Batch-solve $L_k \epsilon_k = z - \mu_k$	$O(KD^2)$
3	Conditional means and variances from $\epsilon_{1:K}$	$O(KD^2)$
4	Weighted CDFs $W \odot F_{d _{\cdot}}(z_d)$	$O(KD)$
5	Sum over components to get $S_{\theta}(z)$	$O(KD)$

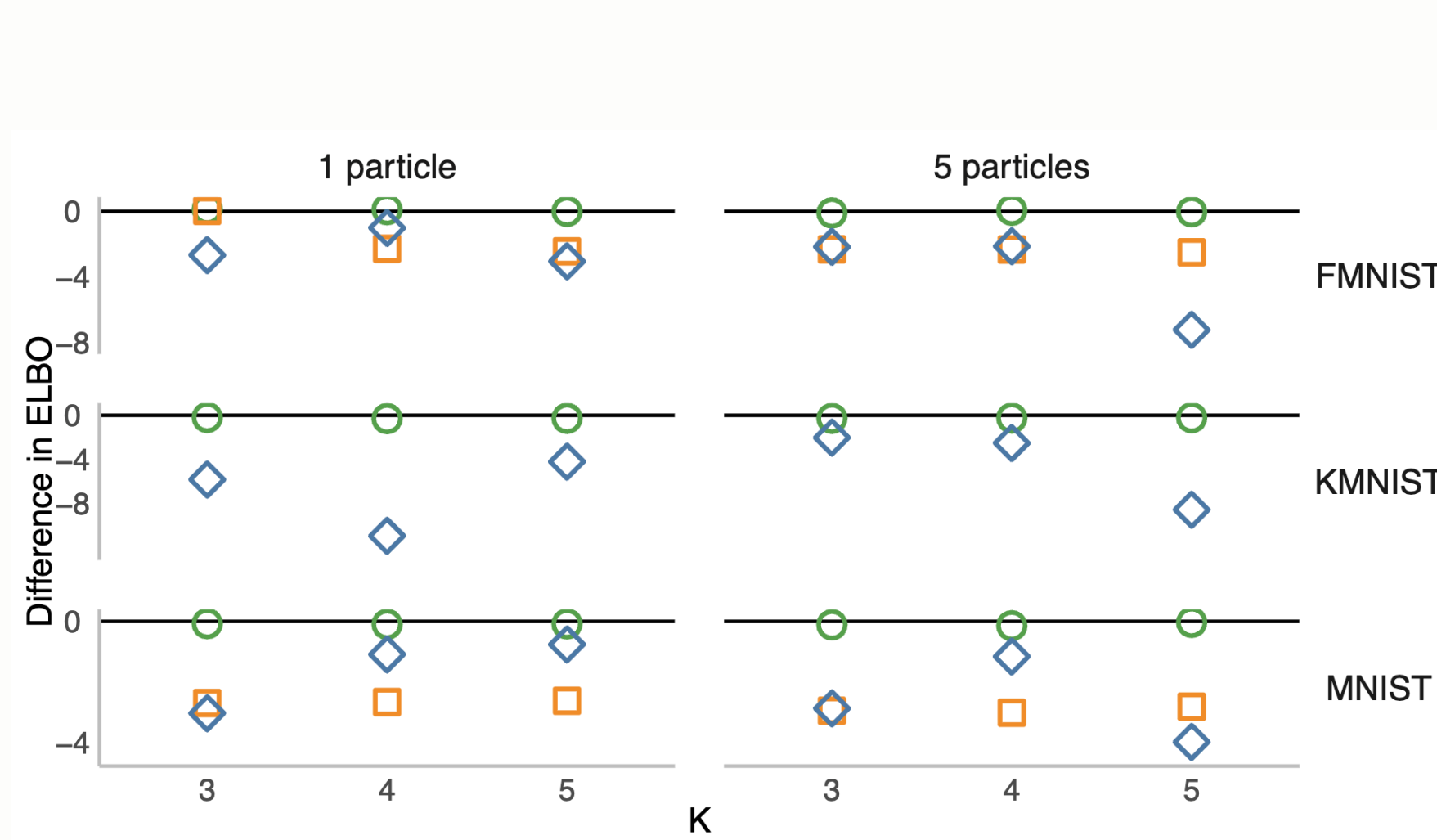
6. Experiments

Baseball · fixed posterior



Test ELBO vs. iteration, $K \in \{5, 21\}$. Stratified converges fastest; post-stratified and J&K are close; implicit lags and runs slowest.

VAEs · MNIST-like, $D = 50$



ELBO minus stratified. Post-stratified **matches or beats** stratified in every setting; J&K diverges on KMNIST.

Chains of mixtures · DMM

Bach chorales; stratified would need $> K^{60}$ evaluations. Test ELBO:

K	Implicit	J&K	Post-st.	Score
2	-7.75	-7.72	-10.64	-11.26
3	-7.81	-7.75	-11.09	-11.25
5	-7.83	-7.85	-11.20	-11.24

Weight-gradient variance dominates here: J&K and implicit win.