Aalto University
School of Science
and Technology

Challenges and Opportunities in High-dimensional Variational Inference

Akash Kumar Dhaka*, Alejandro Catalina*, Manushi Welandawe, Jonathan Huggins, Michael Andersen, Aki Vehtari

Department of Computer Science
Aalto University, School of Science and Technology

December 4, 2021

Background and Motivation

- **Approximating family**, **divergence measure** and **gradient estimators** and their interplay play a key role in **variational inference**
- The complexity of these interactions is aggravated for **high dimensional posteriors**
- These components become even more critical when the goal is to obtain **accurate summaries of the posterior** itself
- The **density ratio** and its **evaluation with MC draws** is the key object of interest

Background and Motivation

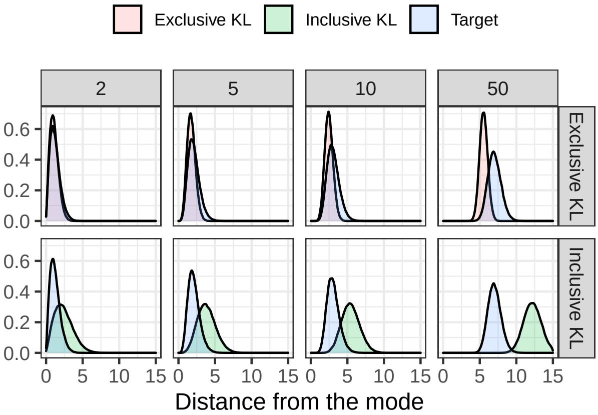


Figure: Distance from the mode for draws of target and approximations for different dimensions $D = [2, 5, 10, 50]$.

Background and Motivation

- When the density ratio is heavy-tailed, even unbiased estimators show a large bias (and large variance) in practice.
- The density ratio is typically heavy-tailed when the typical sets of the target and the approximation do not match.
- For commonly used sample size, the Monte Carlo average is lower than true value with a high probability.
- In higher dimensions, even over-dispersed distributions miss the typical set producing a highly skewed distribution over density ratio.

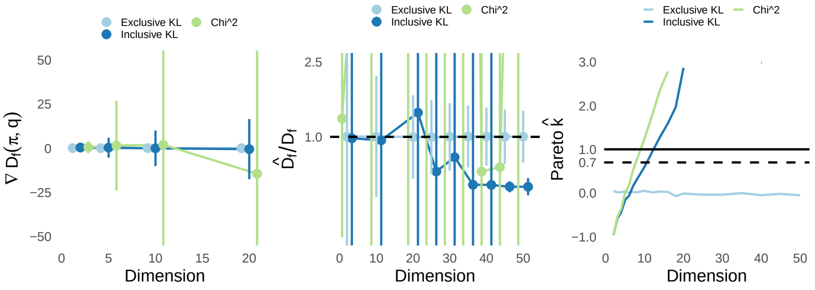
Conceptual framework

- Most common variational divergences can be expressed as a function of the density ratio $w(\theta) = p(\theta, Y)/q(\theta)$ as an f-divergence $D_f(p\|q) = \mathbb{E}_{\theta\sim q}\left[f\left(\frac{p(\theta|Y)}{q(\theta)}\right)\right]$.
- For instance, exclusive KL corresponds to choosing $f = -\log(w)$.
- Reliable BBVI depends on the behavior of $w(\theta)$ since
  1. accurate optimization requires low-variance and (nearly) unbiased gradient estimates $\widehat{G}(\lambda)$,
  2. the quality of variational approximations requires accurate estimates $\widehat{L}(\lambda)$ of variational divergences.

Conceptual framework

- The tail distribution of $w(\theta)$ is well approximated by a general Pareto distribution with parameter k.
- $\lfloor 1/k \rfloor$ determines the number of finite moments of the distribution.
- We can generalize this to the pre-asymptotic behaviour of the gradient and function estimates $\hat{G}, \hat{L}$.
- Approximating their distributions with a generalized Pareto k distribution tell us about their convergence issues in the pre-asymptotic regime.

Conceptual framework



(a) (b) (c)

Figure: Results for correlated Gaussian targets of dimension $D = 1, \ldots, 50$ using exclusive or inclusive KL, and Chi^2 divergences as the variational objective.

Conceptual framework

1. Estimates and gradients of mode-seeking divergences (in particular exclusive KL divergence with log dependence on w) have lower variance and are less biased than those of mass-covering divergences (in particular α-divergences with $\alpha > 0$, with polynomial dependence on w).
2. The degree of polynomial dependence on w determines how rapidly the bias and variance will increase as approximation accuracy degrades – in particular, in high dimensions.
3. The $\hat{k}$ value can be used to diagnose pre-asymptotic reliability of variational objectives. In particular, the α-divergence with $\alpha > 0$ will become unreliable when $\max(1, \alpha) \times \hat{k} > 0.7$, even if w is bounded (by a very large constant).

Experiments on robust regression

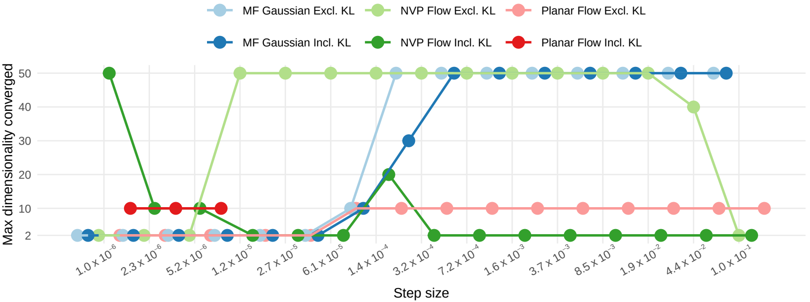
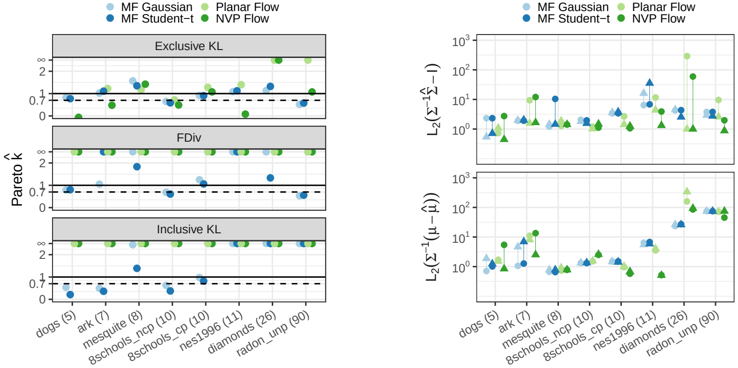


Figure: Maximum dimensionality converged per step size for the robust regression model.

Experiments on real world datasets



(a) (b)

Figure: Results for `posteriordb` experiments. **(a)** Pareto $\hat{k}$ values for BBVI approximations. **(b)** Relative error of mean and covariance estimates for BBVI using exclusive KL (circles) and after PSIS correction (triangles).

Conclusions

- Pareto *k* can be used as a reliable diagnostic that indicates convergence issues/bad approximation.
- Mode-seeking divergences are in practice more stable to optimize and lead to more reasonable results
- Mass-covering divergences do well in low dimensional settings, but are too unstable for higher dimensional targets
- PSIS correction improves the estimation of many quantities of interest, i.e. posterior summaries