4th Symposium on Advances in Approximate Bayesian Inference, 2021 1–11

U-Statistics for Importance-Weighted Variational Inference

Javier Burroni JBURRONI@CS.UMASS.EDU
University of Massachusetts Amherst

Kenta Takatsu* KTAKATSU@ANDREW.CMU.EDU
Carnegie Mellon University

Justin Domke DOMKE@CS.UMASS.EDU and
Daniel Sheldon SHELDON@CS.UMASS.EDU
University of Massachusetts Amherst

Abstract

We propose the use of U-statistics to estimate the (gradients of) the importance-weighted evidence lower bound (IW-ELBO), a variational objective that uses multiple samples from a proposal distribution to lower-bound the log-likelihood. We propose a *complete U-statistic* estimator, which has variance that is never higher than the standard IW-ELBO estimator, and, under certain conditions, the lowest variance of any unbiased estimator. However, it requires evaluating the objective on a large number of subsets of samples from the proposal distribution, which can be computationally expensive. We propose to use incomplete U-statistics as practical alternatives. We find empirically that both methods reduce estimator variance for its gradients with little computational cost, and lead to faster optimization.

1. INTRODUCTION

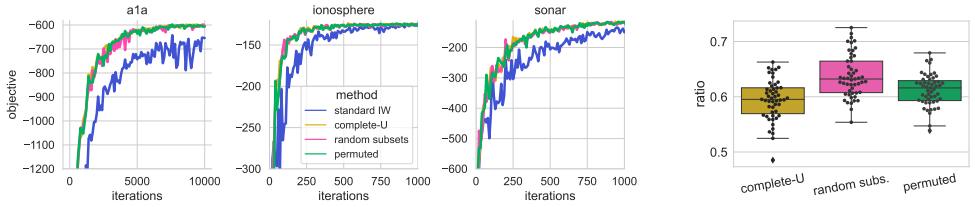
An important recent development in variational inference (VI) is the use of ideas from Monte Carlo sampling to obtain tighter variational bounds (Burda et al., 2016; Maddison et al., 2017; Le et al., 2018; Naesseth et al., 2018; Domke and Sheldon, 2019). These lead to better approximations of the log-likelihood for learning (Burda et al., 2016) and better approximate posterior distributions (Cremer et al., 2017; Domke and Sheldon, 2018; Naesseth et al., 2018; Domke and Sheldon, 2019).

Assume a target distribution $p(z, x) = p(z)p(x \mid z)$ where x is observed and z is latent. VI uses the following evidence lower bound (ELBO), given approximating distribution q, to approximate $\ln p(x)$ (Saul et al., 1996; Blei et al., 2017): $\mathcal{L} = \mathbb{E}_{Z\sim q}\left[\ln \frac{p(Z,x)}{q(Z)}\right] \le \ln p(x)$.

Burda et al. (2016) first showed that a tighter bound can be obtained by using the average of m importance weights within the logarithm. The importance-weighted ELBO (*IW-ELBO*) is $\mathcal{L}_m = \mathbb{E}_{Z_{1:m}}\left[\ln \frac{1}{m}\sum_{i=1}^m \frac{p(Z_i,x)}{q(Z_i)}\right] \le \ln p(x)$, where the expectation is over $Z_1, \ldots, Z_m$ drawn independently from q. We expect Jensen's inequality to provide a tighter bound because the distribution of this sample average is more concentrated around $p(x)$ than the distribution of one estimate. Indeed, $\mathcal{L}_m \ge \mathcal{L}_{m'}$ for $m > m'$ and approaches $\ln p(x)$ as $m \to \infty$ (Burda et al., 2016).

In practice, the IW-ELBO is estimated by sampling. It is convenient to define the *log-weight* random variables $W_i = \ln p(Z_i) - \ln q(Z_i)$ for $Z_i \sim q$ and rewrite the IW-ELBO as

* Work done while at UMass.



(*a*) Objective as a function of iterations.

(*b*) Ratio of the trace of gradients' covariance.

Figure 1: (a) Envelope across different learning rates of the objective on three Bayesian logistic regression models with $m = 4$. Left `a1a` ($d = 120$), center `ionosphere` ($d = 34$), and right `sonar` ($d = 60$). (b) Ratio of the trace of the gradients' covariance when using the methods to that of using standard-IW on `a1a`.

$\mathcal{L}_m = \mathbb{E}_{W_{1:m}}\left[\ln \frac{1}{m}\sum_{i=1}^{m} e^{W_i}\right]$. Then, for example, with $m = 2$, we can draw two log-weights W_1 and W_2 and estimate the IW-ELBO as $\hat{\mathcal{L}}_2 = \ln(e^{W_1} + e^{W_2}) - \ln 2$.
It is standard to use multiple replicates to reduce variance (Rainforth et al., 2018; Domke and Sheldon, 2018). For example, we could draw four log-weights instead of two and estimate the IW-ELBO as $\hat{\mathcal{L}}_{4,2} = \frac{1}{2}\ln(e^{W_1} + e^{W_2}) + \frac{1}{2}\ln(e^{W_3} + e^{W_4}) - \ln 2$. However, using the same n log-weights, we could instead compute

$$\hat{\mathcal{L}}^U_{4,2} = \tfrac{1}{6}\ln(e^{W_1} + e^{W_2}) + \tfrac{1}{6}\ln(e^{W_1} + e^{W_3}) + \tfrac{1}{6}\ln(e^{W_1} + e^{W_4}) + \tfrac{1}{6}\ln(e^{W_2} + e^{W_3}) + \tfrac{1}{6}\ln(e^{W_2} + e^{W_4}) + \tfrac{1}{6}\ln(e^{W_3} + e^{W_4}) - \ln 2,$$

which uses all possible pairs of log-weights instead of two disjoint pairs. This is an instance of a complete U-statistic, a family of estimators introduce by Hoeffding (1948) following the work of Halmos (1946). It is clear that both estimators are unbiased estimators of $\mathcal{L}_2$. Additionally, it was proven by Halmos that the complete U-statistics (and hence $\hat{\mathcal{L}}^U_{4,2}$) are optimal in the sense of having the smallest variance among all unbiased estimators (of $\mathcal{L}_2$ in our case). These two properties suggest that the complete U-statistics are good candidates for the task of optimizing the IW-ELBO, thus motivating our work.
The contributions of this paper are: i) we propose the *complete U-statistic* IW-ELBO estimator, which is the unbiased estimator of $\mathcal{L}_m$ with the smallest variance [Section 2], and describe how to construct an analogous estimator for gradient estimation that is compatible with different base gradient estimators [Section 2.1]; ii) to reduce the computational burden of evaluating the objective on $\binom{n}{m}$ sets for the complete U-statistic, we construct particular *incomplete U-statistics* (Blom, 1976) and prove they achieve most of the variance reduction of complete U-statstics at a modest cost [Section 3]; iii) we show empirically that our approaches consistently reduce the trace of the gradients' covariance compared to the standard IW-ELBO estimator [Section 4], which leads to faster learning [Fig. 1(*a*)].

2. U-STATISTIC ESTIMATORS

We now formalize the examples above by introducing the standard IW-ELBO estimator and its corresponding *complete U-statistic IW-ELBO estimator*, and applying the theory of U-statistics to relate their variances. The theory of U-statistics was developed in a seminal work by Hoeffding (1948) and extends the theory of unbiased estimation introduced by Halmos (1946). Detailed background can be found in the original works, or see (Lee, 1990; van der Vaart, 2000).

Our goal is to estimate the IW-ELBO $\mathcal{L}_m$ and its gradient with respect to parameters of the log-weight distribution. We focus here on IW-ELBO estimation, but an analogous development holds for reparameterization gradients (see Section 2.1). We consider estimators that are functions of n independent log-weights $\mathbf{W} = (W_1, \ldots, W_n)$, where we assume for convenience that $n = rm$ for an integer number of replicates $r \geq 1$.

Estimator 1 *The standard IW-ELBO estimator is*

$$\hat{\mathcal{L}}_{n,m}(\mathbf{W}) = \frac{1}{r}\sum_{j=0}^{r-1} \ln\Big(\frac{1}{m}\sum_{i=1}^{m} e^{W_{i+rj}}\Big).$$

This estimator partitions the n log-weights into r disjoint sets and computes the average of the unbiased estimates from each set. In contrast, estimators based on U-statistics will use overlapping sets of log-weights. To set this up, we introduce notation for size-m subsets of the indices from 1 to n. Let $[\![n]\!] = \{1, \ldots, n\} \subseteq \mathbb{N}$, let $\binom{A}{m}$ denote the set of all subsets of A with exactly m elements, and, for $\mathbf{s} \in \binom{[\![n]\!]}{m}$, let s_i be the ith smallest index in $\mathbf{s}$.

Estimator 2 *The complete U-statistic IW-ELBO estimator is*

$$\hat{\mathcal{L}}^{U}_{n,m}(\mathbf{W}) = \binom{n}{m}^{-1} \sum_{\mathbf{s}\in\binom{[\![n]\!]}{m}} \ln\Big(\frac{1}{m}\sum_{i=1}^{m} e^{W_{s_i}}\Big).$$

In words, $\hat{\mathcal{L}}^{U}_{n,m}$ takes the average of $\ln\big(\frac{1}{m}\sum_{i=1}^{m} e^{W_{s_i}}\big)$ over all distinct subsets $\mathbf{s}$ of m indices. Contrast this to $\hat{\mathcal{L}}_{n,m}$, which takes the same average but over r disjoint subsets of m indices. It follows from linearity of expectation that both $\hat{\mathcal{L}}^{U}_{n,m}$ and $\hat{\mathcal{L}}_{n,m}$ are unbiased estimators for $\mathcal{L}_m$.

The relationship to general U-statistics studied by Halmos (1946) and Hoeffding (1948) is as follows. Given n random variables $W_1, \ldots, W_n$ drawn iid from a distribution F, and an unbiased estimator of the form $\theta(F) = \mathbb{E}\, h(W_1, \ldots, W_m)$ for a *kernel* h that is symmetric in its $m \leq n$ arguments, the complete U-statistic is $\binom{n}{m}^{-1} \sum_{\mathbf{s}\in\binom{[\![n]\!]}{m}} h(W_{s_1}, \ldots, W_{s_m})$. Our estimator is a complete U-statistic for $h(W_1, \ldots, W_m) = \ln\big(\frac{1}{m}\sum_{i=1}^{m} e^{W_i}\big)$.

A result of Halmos (1946) implies that, subject to certain conditions, $\hat{\mathcal{L}}^{U}_{n,m}$ has the smallest variance of any unbiased estimator of the IW-ELBO. The technical conditions are needed to define the class of "unbiased estimators" as ones that are unbiased for all log-weight distributions in a non-trivial class.

We observe in practice that there is a gap between the two variances that leads to practical gains for the complete U-statistic estimator in real VI problems.

2.1. Gradients

Variance reduction via U-statistics also applies to gradient estimators. For example, assume the log-weights are reparameterizable as $W_i = W(\epsilon_i, \theta)$ for ϵ_i drawn iid from a fixed base distribution. By linearity, the reparameterization partial derivative estimator $\frac{\partial}{\partial \theta_j} \hat{\mathcal{L}}^U_{n,m}\big(W(\epsilon_1,\theta),\ldots,W(\epsilon_n,\theta)\big)$ is itself a complete U-statistic corresponding to the kernel $h_j(\epsilon_1,\ldots,\epsilon_m;\theta) = \frac{\partial}{\partial \theta_j} \ln \frac{1}{m} \sum_{i=1}^m e^{W(\epsilon_i,\theta)}$, which is "base" reparameterization gradient estimator. More broadly, we can form a complete U-statistic gradient estimator $\binom{n}{m}^{-1} \sum_{\mathbf{s}\in\binom{[\![n]\!]}{m}} h(\epsilon_{s_1},\ldots,\epsilon_{s_m};\theta)$ from any base gradient estimator, for example, the doubly reparameterized gradient estimator $h_j(\epsilon_1,\ldots,\epsilon_m;\theta) = \sum_{i=1}^m \left(\frac{e^{W(\epsilon_i,\theta)}}{\sum_{k=1}^m e^{W(\epsilon_k,\theta)}}\right)^2 \frac{\partial W(\epsilon_i,\theta)}{\partial \theta_j}$ (Tucker et al., 2018), or the score function estimator.

Because the theory of this section derived from general U-statistic theory, for each base estimator, it holds an exact analog of Prop 9 in the Appendix, this time using h_j. The complete U-statistic estimator will have variance that is never higher, and usually lower, than the base estimator averaged over r disjoint samples. As a concrete result, this implies that $\mathbb{E}\,\|\nabla_\theta \hat{\mathcal{L}}^U_{n,m}\|_2^2 \leq \mathbb{E}\,\|\nabla_\theta \hat{\mathcal{L}}_{n,m}\|_2^2$. The same situation will apply in the next section: results about variance of U-statistic estimators for the IW-ELBO also apply to U-statistic estimators of gradients.

Analogs of Proposition 7 also hold for gradient estimation, but should be taken lightly. Gradient estimators often use properties of the underlying sampling distribution of the log-weights—e.g., that the distribution can be reparameterized, or that $\log p(Z;\theta)$ is differentiable with respect to θ (for the score function estimator)—that make the conditions unlikely to hold. Indeed, it is precisely by using these properties that different approaches are often able to reduce variance.

2.2. Computational Complexity

There are two main factors to consider for the computational complexity of an IW-ELBO estimator:

1) The cost to compute n log-weights $W_i = \ln p(Z_i, x) - \ln q(Z_i)$ for $i \in [\![n]\!]$, and
2) the cost to compute the estimator given the log-weights.

A problem with $\hat{\mathcal{L}}^U_{n,m}$ is that it averages $\binom{n}{m}$ distinct subsets of indices in 2), which is expensive. It should be noted that the individual operations to compute $\hat{\mathcal{L}}^U_{n,m}$ are very simple, while, for many probabilistic models, computing each log-weight is expensive, so, for modest m and n, the computation may still be dominated by Step 1). However, for large enough m and n, Step 2) is impractical.

3. INCOMPLETE U-STATISTIC ESTIMATORS

In practice, we can achieve most of the variance reduction of the complete U-statistic IW-ELBO estimator with only modest computational cost by averaging over only $k \ll \binom{n}{m}$ subsets of indices selected in some way. Such an estimator is called an *incomplete U-statistic*. Incomplete U-statistics were introduced and studied by Blom (1976). Note that the standard IW-ELBO estimator $\hat{\mathcal{L}}_{n,m}$ is itself an incomplete U-statistic, where the $k = r = \frac{n}{m}$ index sets are disjoint. We can improve on this by selecting $k > r$ sets.

Given $\mathcal{S}$ a collection of size-m subsets of $[\![n]\!]$, a general incomplete U-statistic for the IW-ELBO is defined as $\hat{\mathcal{L}}_{\mathcal{S}}(\mathbf{W}) = \frac{1}{|\mathcal{S}|}\sum_{\mathbf{s}\in\mathcal{S}} \ln\Big(\frac{1}{m}\sum_{i=1}^{m} e^{W_{s_i}}\Big)$. We will allow $\mathcal{S}$ to be a multi-set, that is, the same subset may appear more than once.

Estimator 3 (Random subsets) *The* random-subset incomplete-U-statistic estimator for the IW-ELBO *is the estimator* $\hat{\mathcal{L}}_{\mathcal{S}_k}$ *where* $\mathcal{S}_k$ *is a set of* k *subsets* $(\mathbf{s}_i)_{i=1}^k$ *drawn uniformly at random (with replacement) from* $\binom{[\![n]\!]}{m}$.

Estimator 4 (Permuted block) *The* permuted-block incomplete-U-statistic estimator for the IW-ELBO *is the estimator* $\hat{\mathcal{L}}_{\mathcal{S}^\ell_\Pi}$ *with the collection* S^ℓ_Π *defined as follows. Let* π *denote a permutation of* $[\![n]\!]$. *Define* $\mathcal{S}_\pi$ *as the collection obtained by permuting indices according to* π *and then dividing them into* r *disjoint sets of size* m. *That is,*

$$\mathcal{S}_\pi = \Big\{\big\{\pi(1),\pi(2),\dots,\pi(m)\big\},\dots,\big\{\pi\big((r-1)m+1\big),\dots,\pi(rm)\big\}\Big\}$$

Now, let $\mathcal{S}^\ell_\Pi = \biguplus_{\pi\in\Pi}\mathcal{S}_\pi$ *where* Π *is a collection of* ℓ *random permutations and* $\biguplus$ *denotes union as a multiset. The total number of sets in* $\mathcal{S}^\ell_\Pi$ *is* $k = r\ell$.

Both incomplete-U-statistic estimators can achieve variance reduction in practice for a large enough number of sets k, but the permuted block estimator has an advantage: its variance with k subsets is never more than that of the random subset estimator with k subsets, and never more than the variance of the standard IW-ELBO estimator (and usually smaller). On the other hand, the variance of the random subset estimator is more than that of the standard estimator unless $k \geq k_0$ for some threshold $k_0 > r$.

Proposition 5 *Given* m *and* $n = rm$, *the variances of the estimators satisfy the following partial ordering:*

$$\underbrace{\mathrm{Var}[\hat{\mathcal{L}}^U_{n,m}]}_{\text{complete}} \overset{\text{(a)}}{\leq} \underbrace{\mathrm{Var}[\hat{\mathcal{L}}_{\mathcal{S}^\ell_\Pi}]}_{\text{permuted}} \begin{array}{l} \overset{\text{(b)}}{\leq} \overbrace{\mathrm{Var}[\hat{\mathcal{L}}_{n,m}]}^{\text{standard}} \\ \overset{\text{(c)}}{\leq} \underbrace{\mathrm{Var}[\hat{\mathcal{L}}_{\mathcal{S}_{r\ell}}]}_{\text{random subset}} \end{array}. \tag{1}$$

Moreover, if $\ell > 1$ *and* $\mathrm{Var}[\hat{\mathcal{L}}^U_{n,m}] < \mathrm{Var}[\hat{\mathcal{L}}_{n,m}]$, *then* (a) *is strict; if* $r > 1$, *then* (b) *is strict.* (Note that $\hat{\mathcal{L}}_{\mathcal{S}^\ell_\Pi}$ and $\hat{\mathcal{L}}_{\mathcal{S}_{r\ell}}$ both use $k = r\ell$ subsets.)

A remarkable property of the permuted-block estimator is that we can choose the number of permutations ℓ based on how much of the available variance reduction we want to achieve. Say we would like to achieve 95% of the variance reduction; then it suffices to set $\ell = 20$. The following Proposition formalizes this result.

Proposition 6 *Given* m *and* $n = rm$, *for* $\ell \in \mathbb{N}$ *the permuted-block estimator achieves a* $(1 - 1/\ell)$ *fraction of the variance reduction provided by the complete U-statistic IW-ELBO estimator, i.e.,*

$$\underbrace{\mathrm{Var}[\hat{\mathcal{L}}_{n,m}]}_{\text{standard}} - \underbrace{\mathrm{Var}[\hat{\mathcal{L}}_{\mathcal{S}^\ell_\Pi}]}_{\text{permuted}} = (1-\tfrac{1}{\ell})\big(\underbrace{\mathrm{Var}[\hat{\mathcal{L}}_{n,m}]}_{\text{standard}} - \underbrace{\mathrm{Var}[\hat{\mathcal{L}}^U_{n,m}]}_{\text{complete}}\big).$$

4. EXPERIMENTS

In this section we analyze empirically the methods proposed in this paper for IW-VI (i.e., optimizing the IW-ELBO with respect to the parameters of the proposal distribution q). The general setup is shared across experiments: we experiment with Bayesian logistic regression using a diagonal Gaussian prior and a posterior approximated with a Gaussian distribution with full covariance matrix.[1] The models were optimized using stochastic gradient descent with fixed learning rate. We will set $n = 16$, except for the analysis of the required time [see below]. We ran the experiments with $m \in [2, 4, 8]$ and we report the results for $m = 4$. We set $\ell = 20$ for the permuted block estimator, and for the random subsets estimator, we put $k = 20\frac{n}{m}$. We show the results on the `a1a` ($d = 120$), `ionosphere` ($d = 34$) and `sonar` ($d = 60$) datasets, but additional figures can be found in the Appendix. Ultimately, our goal is to provide a more efficient optimization method. To empirically show this, we optimized each alternative over a range of different learning rates.[2] In Fig. 1(a) we show the envelope of the objective for all methods, i.e., at every iteration and method, we take the maximum value of the objective across all learning rates. The Figure shows that the presented methods outperform the standard IW-ELBO estimator. The reason is that those methods allow for higher learning rates.

We conjecture that this phenomenon is due to the total variance of the gradients, i.e., the trace of the covariance matrix of the gradients. We estimated this quantity for each estimator taking 200 samples every 200 iterations. We present in Fig. 1(b) the ratio to the estimation of the total variance of the gradients for the standard IW-ELBO estimator using the `a1a` dataset. The ratio is approximately 60% for all methods, with the random subsets estimator showing the highest ratio, and the complete U-statistic the lowest. Moreover, the permuted estimator with $\ell = 20$ achieves approximately 95.58% of the variance reduction, as predicted by Proposition 6.

To complete, we provide in Table 1 the times required to complete $1k$ iterations of the optimization. Here we used $n = 24$ and $m = 12$, which makes it a challenging setting for the complete U-statistic IW-ELBO estimator because there are $\binom{24}{12} = 2,704,156$ sets. As expected, the only outlier method is the complete U-statistic. The differences in the required times by the other methods to the time required by the standard IW-ELBO estimator are not statistically significant and can be attributed to noise in the environment.

1. That is, $p(\theta) = \mathcal{N}(\theta; \mathbf{0}, \sigma^2\mathbf{I}_d)$ and $p(y \mid \theta) = \prod_{i=1}^{N} \text{Bernoulli}\big(y_i; \text{logistic}(\theta^T x_i)\big)$ for fixed $x_i \in \mathbb{R}^d$, and $W = \ln p(\theta, y) - \ln q(\theta)$ for $\theta \sim q(\theta)$ with $q(\theta) = \mathcal{N}(\theta; \mu, LL^T)$; we optimize over (μ, L), where L is constrained to be lower triangular with positive diagonal.
2. We used 15 logarithmically-spaced learning rates.

Table 1: Times for 1000 iterations on the `a1a` dataset ($d = 120$) with $n = 24$, $m = 12$, and 15 rep. of the experiments.

Method	Time (s)	
	Mean	Std
$\hat{\mathcal{L}}_{24,12}$ standard IW-ELBO	2.22	0.45
$\hat{\mathcal{L}}^{U}_{24,12}$ complete U	2265.48	43.25
$\hat{\mathcal{L}}_{\mathcal{S}_{20\frac{24}{12}}}$ random subsets	2.07	0.11
$\hat{\mathcal{L}}^{20}_{\mathcal{S}_{\Pi}}$ permuted block	2.16	0.35

5. Citations and Bibliography

Acknowledgments

Acknowledgements JB would like to thank Tomás Geffner for earlier discussions about this project.
This material is based upon work supported by the National Science Foundation under Grant No. 1908577.

References

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Gunnar Blom. Some properties of incomplete U-statistics. *Biometrika*, 63(3):573–580, 1976.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance Weighted Autoencoders. In *ICLR*, 2016.

Chris Cremer, Quaid Morris, and David Duvenaud. Reinterpreting importance-weighted autoencoders. *arXiv preprint arXiv:1704.02916*, 2017.

Justin Domke and Daniel Sheldon. Importance weighting and variational inference. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 4475–4484, 2018.

Justin Domke and Daniel R. Sheldon. Divide and couple: Using Monte Carlo variational objectives for posterior approximation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 338–347, 2019.

Paul R Halmos. The theory of unbiased estimation. *The Annals of Mathematical Statistics*, 17(1):34–43, 1946.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19:273–325, 1948.

Tuan Anh Le, Maximilian Igl, Tom Rainforth, Tom Jin, and Frank Wood. Auto-encoding sequential Monte Carlo. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

Laurence Lock Lee. *U-statistics. Theory and Practice.* CRC Press, 1990. ISBN 9781351405867.

Chris J. Maddison, Dieterich Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, and Yee Whye Teh. Filtering variational objectives. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6573–6583, 2017.

Christian A. Naesseth, Scott W. Linderman, Rajesh Ranganath, and David M. Blei. Variational sequential Monte Carlo. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, volume 84 of *Proceedings of Machine Learning Research*, pages 968–977. PMLR, 2018.

Tom Rainforth, Adam Kosiorek, Tuan Anh Le, Chris Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter variational bounds are not necessarily better. In *International Conference on Machine Learning*, pages 4277–4285. PMLR, 2018.

Lawrence K Saul, Tommi Jaakkola, and Michael I Jordan. Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4:61–76, 1996.

George Tucker, Dieterich Lawson, Shixiang Gu, and Chris J Maddison. Doubly reparameterized gradient estimators for Monte Carlo objectives. In *International Conference on Learning Representations*, 2018.

Aad W van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, 2000.

Appendix A. Additional results

Proposition 7 *Let $\mathbb{E}_F[\,\cdot\,]$ and $\mathrm{Var}_F[\,\cdot\,]$ denote expectation and variance with respect to log-weights $W_1, \ldots, W_n$ drawn independently from distribution F, and denote the IW-ELBO with log-weight distribution F by $\mathcal{L}_m(F) = \mathbb{E}_F\big[\ln \frac{1}{m}\big(\sum_{i=1}^m e^{W_i}\big)\big]$. Let $\mathcal{F}$ denote the set of distributions supported on a finite subset of $\mathbb{R}$. Suppose Φ is any estimator such that $\mathbb{E}_F[\Phi(W_{1:n})] = \mathcal{L}_m(F)$ for all $F \in \mathcal{F}$. Then,*

$$\mathrm{Var}_F[\hat{\mathcal{L}}^U_{n,m}(W_{1:n})] \leq \mathrm{Var}_F[\Phi(W_{1:n})]$$

whenever the latter quantity is defined, for any *distribution F on the real numbers (up to conditions of measurability and integrability).*

Proof The result is a direct application of Theorem 5 of Halmos (1946). ■

For IW-ELBO estimation, the conditions are rather mild: we expect an IW-ELBO estimator to work for generic log-weight distributions. For gradient estimation, we take the conclusion lightly, because gradient estimators often use specific properties of the underlying distributions; see Section 2.1.

A.1. Variance Comparison

How much variance reduction is possible with the complete U-statistic IW-ELBO estimator? This section shows the variance of the $\hat{\mathcal{L}}^U_{n,m}$ is never more than that of $\hat{\mathcal{L}}_{n,m}$, and is strictly less under certain conditions (that occur in practice), using classical bounds on U-statistic variance due to Hoeffding (1948). Since $\hat{\mathcal{L}}^U_{n,m}$ is a sum of terms, one for each $\mathbf{s} \in \binom{[\![n]\!]}{m}$, its variance depends on the covariances between pairs of terms for index sets $\mathbf{s}$ and $\mathbf{s}'$, which in turn depends on how many indices are shared by $\mathbf{s}$ and $\mathbf{s}'$. This motivates the following definition:

Definition 8 *Let $W_1, \ldots, W_{2m}$ be* i.i.d. *log-weights. For $0 \leq c \leq m$, take $\mathbf{s}, \mathbf{s}' \in \binom{[\![n]\!]}{m}$ with $|\mathbf{s} \cap \mathbf{s}'| = c$. Define*

$$\zeta_c = \mathrm{Cov}\Big[\ln\Big(\sum_{i=1}^m \tfrac{1}{m} e^{W_{s_i}}\Big), \ln\Big(\sum_{i=1}^m \tfrac{1}{m} e^{W_{s'_i}}\Big)\Big],$$

which does not depend on the particular choices of $\mathbf{s}$ and $\mathbf{s}'$.

For example, when $m = 2$ we have $\zeta_0 = 0$,

$$\zeta_1 = \mathrm{Cov}[\ln(\tfrac{1}{2}e^{W_1} + \tfrac{1}{2}e^{W_2}), \ln(\tfrac{1}{2}e^{W_1} + \tfrac{1}{2}e^{W_3})], \qquad \text{and} \qquad \zeta_2 = \mathrm{Var}[\ln(\tfrac{1}{2}e^{W_1} + \tfrac{1}{2}e^{W_2})].$$

Then, due to Hoeffding's classical result,

Proposition 9 *Let $\hat{\mathcal{L}}_{n,m}$ and $\hat{\mathcal{L}}^U_{n,m}$ be as in* Estimators 1 *and* 2 *for $n = rm$ with $r \in \mathbb{N}$. Then*

$$\tfrac{m^2}{n}\zeta_1 \leq \mathrm{Var}[\hat{\mathcal{L}}^U_{n,m}] \leq \tfrac{m}{n}\zeta_m = \mathrm{Var}[\hat{\mathcal{L}}_{n,m}].$$

Moreover, for a fixed m, the quantity $n\,\mathrm{Var}[\hat{\mathcal{L}}^U_{n,m}]$ tends to its lower bound $m^2\zeta_1$ as n increases.

Proof The inequalities and asymptotic statement follow directly from Theorem 5.2 of Hoeffding (1948). The equality follows from the definition of ζ_m. ■

We can now proceed to prove Proposition 5 and 6 from the paper.

Proposition 5 *Given m and $n = rm$, the variances of the estimators satisfy the following partial ordering:*

$$\mathrm{Var}[\hat{\mathcal{L}}^U_{n,m}] \leq \mathrm{Var}[\hat{\mathcal{L}}_{\mathcal{S}^\ell_\Pi}] \underset{\text{(b)}}{\overset{\text{(a)}}{\leq}} \mathrm{Var}[\hat{\mathcal{L}}_{n,m}] \leq \mathrm{Var}[\hat{\mathcal{L}}_{\mathcal{S}_{r\ell}}]. \tag{1}$$

Moreover, if $\ell > 1$ and $\mathrm{Var}[\hat{\mathcal{L}}^U_{n,m}] < \mathrm{Var}[\hat{\mathcal{L}}_{n,m}]$, then (a) *is strict; if $r > 1$, then* (b) *is strict.* (Note that $\hat{\mathcal{L}}_{\mathcal{S}^\ell_\Pi}$ and $\hat{\mathcal{L}}_{\mathcal{S}_{r\ell}}$ both use $k = r\ell$ subsets.)

Proof The complete U-statistic is at the lhs of (1) because all the estimators are unbiased, and $\hat{\mathcal{L}}^U_{n,m}$ is the one with smallest variance [cf. Prop. 7].
By Def. 8, if $\mathbf{s}$ and $\mathbf{s}'$ are uniformly drawn from $\binom{[\![n]\!]}{m}$, we have

$$\mathbb{E}[\zeta_{|\mathbf{s}\cap\mathbf{s}'|}] = \mathrm{Var}[\hat{\mathcal{L}}^U_{n,m}]. \tag{2}$$

Observe now that, from the definition of $\hat{\mathcal{L}}_{\mathcal{S}^\ell_\Pi}$ and for a given permutation π, all sets in $\mathcal{S}_\pi$ will be independent. Hence, all dependencies between different sets are due to relations between permutations, i.e., each of the ℓr terms will have a dependency with the $(\ell-1)r$ terms not in the same permutation. Therefore, it follows from (2) that the total variance of $\hat{\mathcal{L}}_{\mathcal{S}^\ell_\Pi}$ is

$$\mathrm{Var}[\hat{\mathcal{L}}_{\mathcal{S}^\ell_\Pi}] = \tfrac{1}{\ell r}\zeta_m + (1 - \tfrac{1}{\ell})\,\mathrm{Var}[\hat{\mathcal{L}}^U_{n,m}], \tag{3}$$

i.e., a convex combination of $\frac{1}{r}\zeta_m = \mathrm{Var}[\hat{\mathcal{L}}_{n,m}]$ and $\mathrm{Var}[\hat{\mathcal{L}}^U_{n,m}]$. Hence, (a) holds.
By a similar argument, the total variance of $\hat{\mathcal{L}}_{\mathcal{S}_{r\ell}}$ is

$$\mathrm{Var}[\hat{\mathcal{L}}_{\mathcal{S}_{r\ell}}] = \tfrac{1}{\ell r}\zeta_m + (1 - \tfrac{1}{r\ell})\,\mathrm{Var}[\hat{\mathcal{L}}^U_{n,m}].$$

Then, (b) holds because

$$\mathrm{Var}[\hat{\mathcal{L}}_{\mathcal{S}^\ell_\Pi}] - \mathrm{Var}[\hat{\mathcal{L}}_{\mathcal{S}_{r\ell}}] = \tfrac{1}{\ell}(\tfrac{1}{r} - 1)\,\mathrm{Var}[\hat{\mathcal{L}}^U_{n,m}] \leq 0.$$

■

Proposition 6 *Given m and $n = rm$, for $\ell \in \mathbb{N}$ the permuted-block estimator achieves a $(1 - 1/\ell)$ fraction of the variance reduction provided by the complete U-statistic IW-ELBO estimator, i.e.,*

$$\mathrm{Var}[\hat{\mathcal{L}}_{n,m}] - \mathrm{Var}[\hat{\mathcal{L}}_{\mathcal{S}^\ell_\Pi}] = (1 - \tfrac{1}{\ell})(\mathrm{Var}[\hat{\mathcal{L}}_{n,m}] - \mathrm{Var}[\hat{\mathcal{L}}^U_{n,m}]).$$

Proof Indeed, this follows directly from Eq. (3). ■

Appendix B. Experiments

B.1. Variance of Objective

The use of U-statistics was motivated because they provide a framework to potentially reduce the variance of the objective (and gradients) within IW-VI, and we confirm this empirically. We performed IW-VI using the complete U-statistic $\hat{\mathcal{L}}^U_{n,m}$, and every 200 iterations we estimated the variance of the objective for the alternatives presented: the standard IW-ElBO estimator, the complete U-statistic IW-ELBO, the permuted-block estimator with $\ell = 20$, and the random subsets estimator with $k = 20\frac{n}{m}$, i.e., a number of sets equal to the permuted version. Recall that, according to Prop. 6, $\ell = 20$ implies that the permuted estimator achieves a 95% of the variance reduction provided by the complete-U-statistic IW-ELBO. In Figure 2, we show the ratio of the variance of the alternatives to that of the standard IW-ELBO estimator for the `a1a` dataset. This Figure confirms that it is possible to reduce the variance of the objective by considering the U-statistics. Moreover, the estimators can be ordered by their variances, with the complete U-statistic IW-ELBO estimator showing the smallest variance, the permuted-block estimator in a middle ground, and finally the random subsets estimator being the one with the smallest variance reduction.

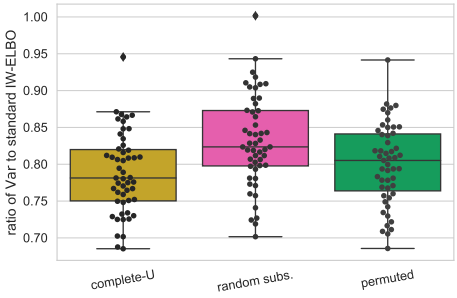


Figure 2: Ratio of the objective's variance when using a given method to that of using standard-IW on `a1a`.

B.2. Additional Figures

In this section we augment the Figures shown in the paper with two additional datasets, `australian` and `mushrooms`, and for $m \in [2, 4, 8]$. In all cases we used $n = 16$ independently sampled log-weights $W_1, \ldots, W_{16}$. We trained all models for $10k$ iterations, but we show results for all iterations only for the largest models, i.e., `a1a` ($d = 120$) and `mushrooms` ($d = 112$), and the rest, i.e, `australian` ($d = 14$), `ionosphere` ($d = 34$) and `sonar` ($d = 60$), we show the results for the first $1k$ iterations.
Additionally, we reported in Fig. 4 the objective achieved by each method after 2k and 10k iterations as a function of the learning rate [cf. Domke and Sheldon (2018)]. In each case, we evaluated the learned approximating distribution using the standard IW-ELBO estimator with $n' = 1000 * m$.

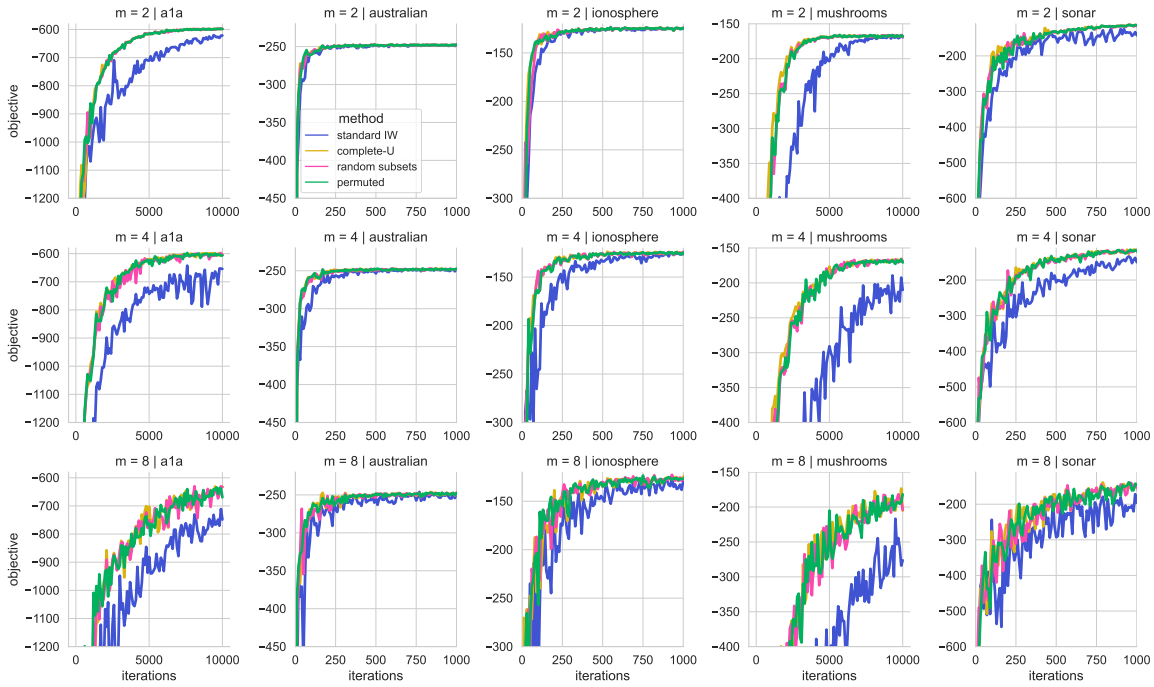


Figure 3: Envelope across different learning rates of the objective on five Bayesian logistic regression models with the methods presented in this paper and $m \in [2, 4, 8]$. From left to right: `a1a` ($d = 120$), `australian` ($d = 14$), `ionosphere` ($d = 34$), `mushrooms` ($d = 112$) and `sonar` ($d = 60$). [See Section 4.]

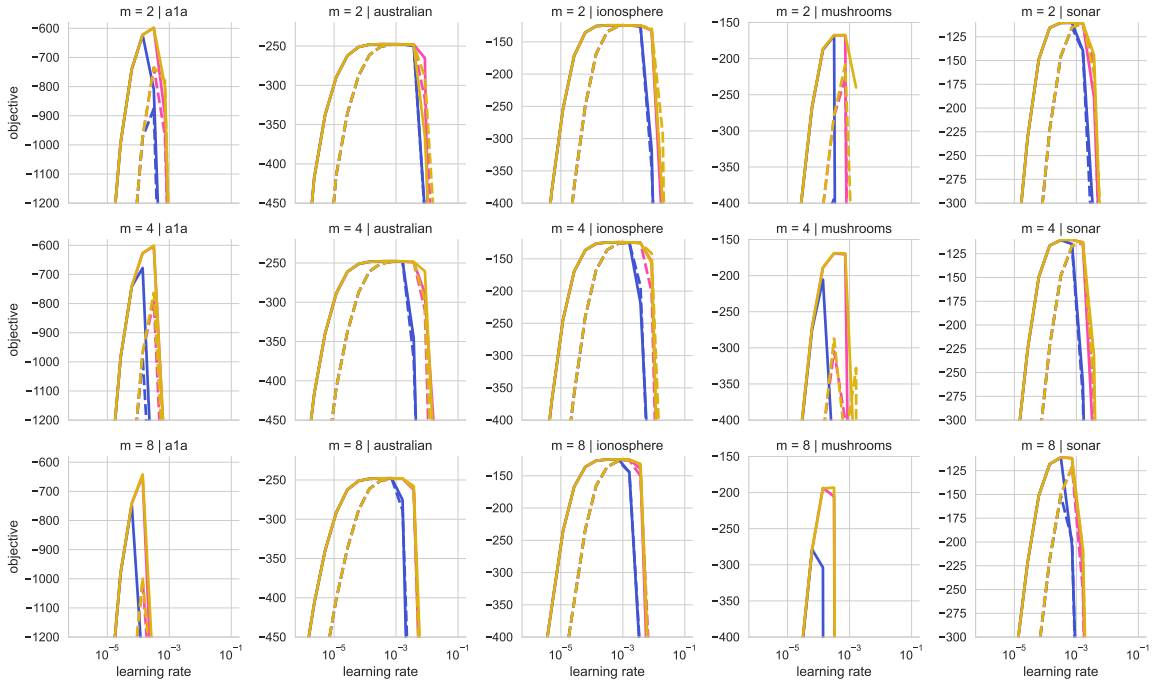


Figure 4: Value of the objective after 2k iterations (dashed lines) and 10k iterations (solid lines) on five Bayesian logistic regression models using $m \in [2, 4, 8]$, and optimized using the standard IW-ELBO, complete-U-statistic and random subsets estimators. Consistently across datasets and m, the alternatives allow to use higher learning rate values.