Divide and Couple: Using Monte Carlo Variational Objectives for Posterior Approximation

Justin Domke and Daniel Sheldon, University of Massachusetts Amherst

- Recent work uses **better estimators** for **better likelihood bounds**.
- But how to apply to "**pure probabilistic**" variational inference (VI)?
- We show that for any unbiased estimator, an **augmented posterior** can be constructed using **couplings**.
- We give a framework of transforming "estimator-coupling-pairs" to easily construct couplings for many Monte Carlo Methods.

1. The Problem

Take some distribution $p(\boldsymbol{z},\boldsymbol{x})$ with $\boldsymbol{x}$ fixed.

Observation: If R is a random variable with $\mathbb{E}\, R = p(\boldsymbol{x})$ then $\mathbb{E}\log R \leq \log p(\boldsymbol{x})$.

Example: Take $R = \frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}$ for z $\sim q$ Gaussian, optimize q:

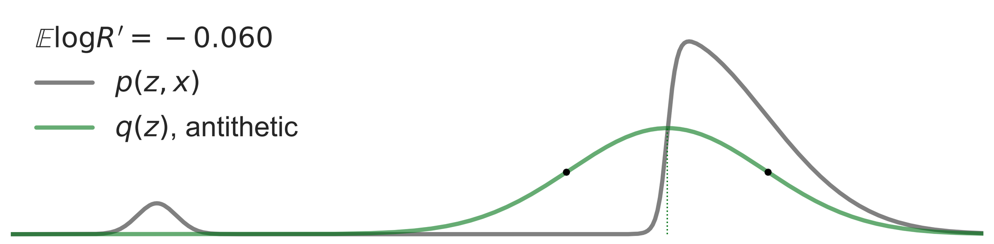
Decomposition: $KL\left(q(\boldsymbol{z}) \| p(\boldsymbol{z}|\boldsymbol{x})\right) = \log p(\boldsymbol{x}) - \mathbb{E}\log R.$

- Likelihood bound: ✓ $\quad(\mathbb{E}\log R \leq \log p(\boldsymbol{x}))$
- Posterior approximation: ✓ $\quad$(q is close to p)

Recent work: Better Monte Carlo estimators R.

Second Example: Antithetic Sampling: Let $T(\boldsymbol{z})$ "flip" z around mean of q.

$$R' = \frac{1}{2}\left(\frac{p(\boldsymbol{z},\boldsymbol{x}) + p(T(\boldsymbol{z}),\boldsymbol{x})}{q(\boldsymbol{z})}\right)$$



- Likelihood bound: ✓ $\quad(\mathbb{E}\log R \leq \log p(\boldsymbol{x}))$
- Posterior approximation: × $\quad$(q is *not* close to p)

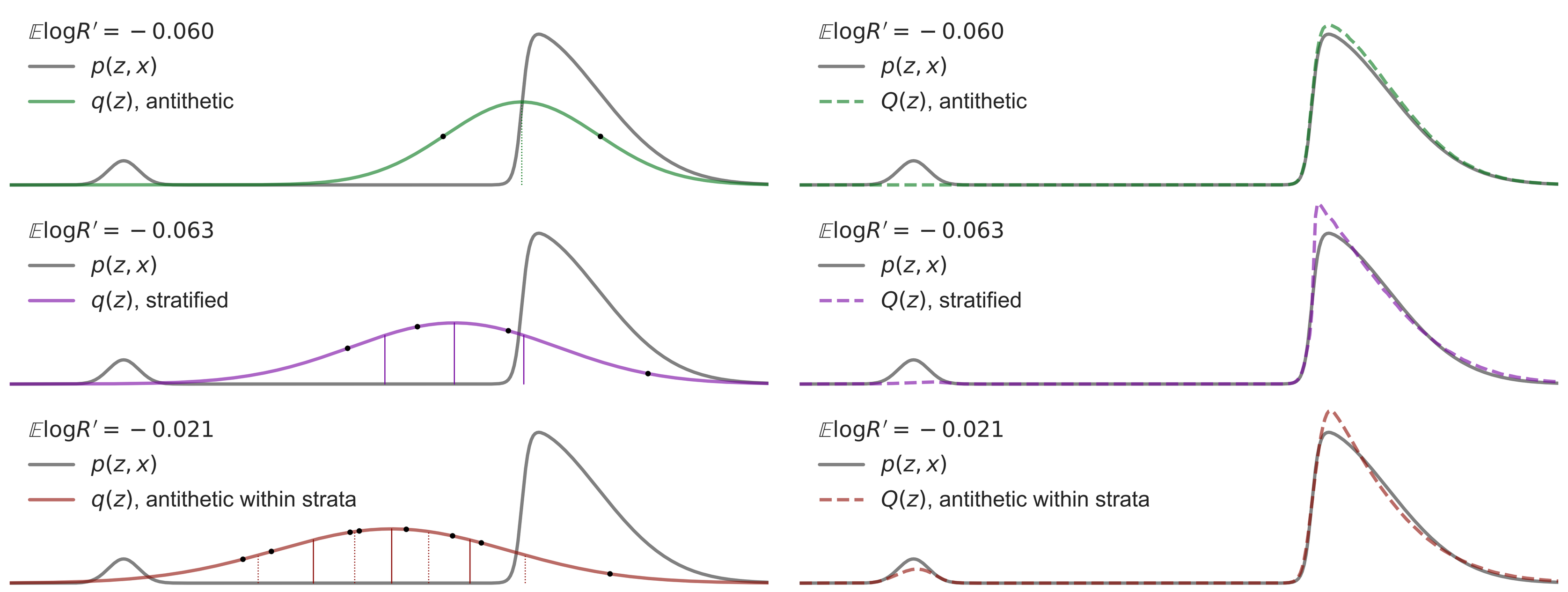
This paper: Is some *other* distribution close to p?

- Felix V. Agakov and David Barber. An Auxiliary Variational Method. NeurIPS, 2004.
- Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. Journal of the Royal Statistical Society: Series B, 72:269–342, 2010.
- Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance Weighted Autoencoders. ICLR, 2015.
- Christian Andersson Naesseth. Machine Learning Using Approximate Inference: Variational and Sequential Monte Carlo Methods. PhD thesis, Linköping University, 2018.
- Tuan Anh Le, Maximilian Igl, Tom Rainforth, Tom Jin, and Frank Wood. Auto-Encoding Sequential Monte Carlo. ICLR, 2018.
- Chris J Maddison, John Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, and Yee Teh. Filtering Variational Objectives. NeurIPS, 2017.

2. Main Framework

Contribution of this paper: Given estimator with $\mathbb{E}\, R = p(\boldsymbol{x})$, we show how to construct $Q(\boldsymbol{z})$ such that

$$KL\left(Q(\boldsymbol{z}) \| p(\boldsymbol{z}|\boldsymbol{x})\right) \leq \log p(\boldsymbol{x}) - \mathbb{E}\log R.$$



2.1 Intuition

- An unbiased estimator $\mathbb{E}_{\omega}\, R(\omega) = p(\boldsymbol{x})$ is not enough!
- We suggest: Need a *coupling*: $\mathbb{E}_{\omega}\, R(\omega) a(z|\omega) = p(\boldsymbol{z},\boldsymbol{x})$.
- Define *augmented* distributions in state space (z, ω).
- Tightening $\log p(\boldsymbol{x}) - \mathbb{E}\log R$ is equivalent to VI on the augmented distributions.

2.2 Divide

1. Random variable R with underlying sample space ω

$$R(\omega),\ \omega \sim Q(\omega).$$

2. Assume unbiased

$$\underset{Q(\omega)}{\mathbb{E}}\left[R(\omega)\right] = p(\boldsymbol{x}).$$

3. Define "extended target"

$$P^{\mathrm{MC}}(\omega,\boldsymbol{x}) = Q(\omega)R(\omega)$$

4. Then

$$KL\left(Q(\omega) \middle\| P^{\mathrm{MC}}(\omega|x)\right) = \log p(\boldsymbol{x}) - \mathbb{E}\log R.$$

Fine... But where is z?

2.3 Couple

2. Assume coupling $a(z|\boldsymbol{\omega})$

$$\underset{Q(\boldsymbol{\omega})}{\mathbb{E}}\left[R(\boldsymbol{\omega})a(\boldsymbol{z}|\boldsymbol{\omega})\right] = p(\boldsymbol{z},\boldsymbol{x}).$$

3. Define
   (a) "extended target"

$$P^{\mathrm{MC}}(\boldsymbol{z},\boldsymbol{\omega},\boldsymbol{x}) = Q(\boldsymbol{\omega})R(\boldsymbol{\omega})a(\boldsymbol{z}|\boldsymbol{\omega},\boldsymbol{x})$$

   (b) "extended proposal"

$$Q^{\mathrm{MC}}(\boldsymbol{z},\boldsymbol{\omega}) = Q(\boldsymbol{\omega})a(\boldsymbol{z}|\boldsymbol{\omega},\boldsymbol{x})$$

4. Then

$$KL\left(Q^{\mathrm{MC}}(\mathrm{z},\omega) \middle\| P^{\mathrm{MC}}(\mathrm{z},\omega|\boldsymbol{x})\right) = \log p(\boldsymbol{x}) - \mathbb{E}\log R.$$

2.4 Summary

- Tightening a bound $\log p(\boldsymbol{x}) - \mathbb{E}\log R$ is equivalent to VI in an *augmented state space* $(\boldsymbol{\omega},\boldsymbol{z})$.
- To sample from $Q(\boldsymbol{z})$ draw $\boldsymbol{\omega} \sim Q(\boldsymbol{\omega})$ then $\boldsymbol{z} \sim a(\boldsymbol{z}|\boldsymbol{\omega})$.
- We give couplings for:
  - Antithetic sampling
  - Stratified sampling
  - Quasi Monte Carlo
  - Latin hypercube sampling
  - Arbitrary recursive combinations of above

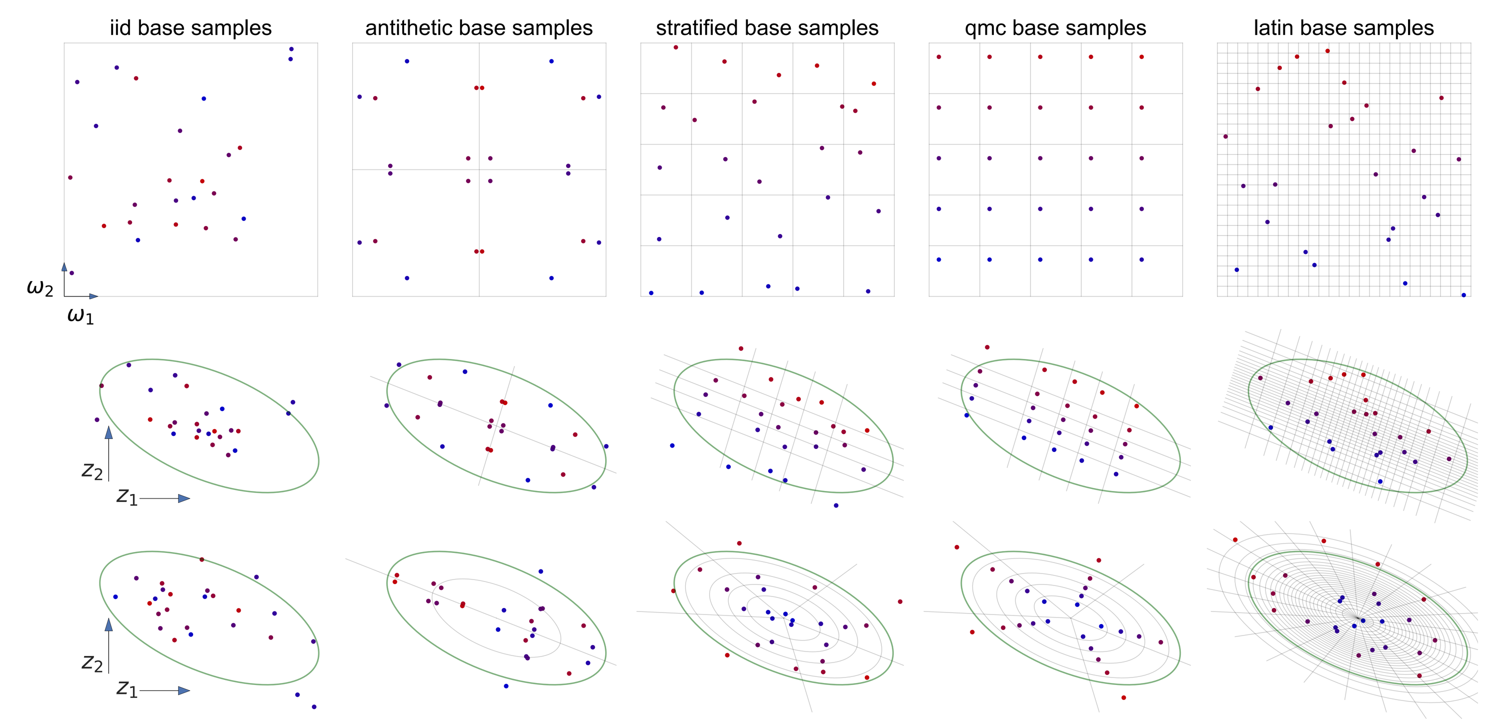
3. Transforming Couplings

Take base estimator-coupling pair (R_0, a_0) transform to create new pair (R, a).

Description	$R(\cdot)$	$a(z\|\cdot)$
IID Mean $\boldsymbol{\omega}_1 \cdots \boldsymbol{\omega}_M \sim Q_0$ i.i.d.	$\frac{1}{M}\sum_{m=1}^{M} R_0(\omega_m)$	$\frac{\sum_{m=1}^{M} R_0(\omega_m)a_0(z\|\omega_m)}{\sum_{m=1}^{M} R_0(\omega_m)}$
Stratified Sampling $\Omega_1 \cdots \Omega_M$ partition Ω, $\boldsymbol{\omega}_1^m \cdots \boldsymbol{\omega}_{M_n}^1 \sim Q_0$ restricted to Ω_m, $\mu_m = Q_0(\boldsymbol{\omega} \in \Omega_m)$.	$\sum_{m=1}^{M} \frac{\mu_m}{N_m} \sum_{n=1}^{N_m} R_0\left(\omega_n^m\right)$	$\frac{\sum_{m=1}^{M} \frac{\mu_m}{N_m} \sum_{n=1}^{N_m} R_0\left(\omega_n^m\right) a_0(z\|\omega_n^m)}{\sum_{m=1}^{M} \frac{\mu_m}{N_m} \sum_{n=1}^{N_m} R_0\left(\omega_n^m\right)}$
Antithetic Sampling $\boldsymbol{\omega} \sim Q_0$. For all m, $T_m(\boldsymbol{\omega}) \overset{d}{=} \boldsymbol{\omega}$.	$\frac{1}{M}\sum_{m=1}^{M} R_0\left(T_m(\omega)\right)$	$\frac{\sum_{m=1}^{M} R_0(T_m(\omega))\, a_0\left(z\|T_m(\omega)\right)}{\sum_{m=1}^{M} R(T_m(\omega))}$
Randomized Quasi Monte Carlo $\boldsymbol{\omega} \sim \mathrm{Unif}([0,1]^d)$, $\bar{\omega}_1, \cdots \bar{\omega}_M$ fixed, $T_m(\omega) = F^{-1}\left(\bar{\omega}_m + \omega \text{ (mod 1)}\right)$	$\frac{1}{M}\sum_{m=1}^{M} R_0\left(T_m(\omega)\right)$	$\frac{\sum_{m=1}^{M} R_0(T_m(\omega))\, a_0\left(z\|T_m(\omega)\right)}{\sum_{m=1}^{M} R_0(T_m(\omega))}$
Latin Hypercube Sampling $\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_M$ jointly sampled from Latin hypercube [21, Ch. 10.3], $T = F^{-1}$.	$\frac{1}{M}\sum_{m=1}^{M} R_0\left(T(\omega_m)\right)$	$\frac{\sum_{m=1}^{M} R_0(T(\omega_m))\, a_0\left(z\|T(\omega_m)\right)}{\sum_{m=1}^{M} R_0(T(\omega_m))}$

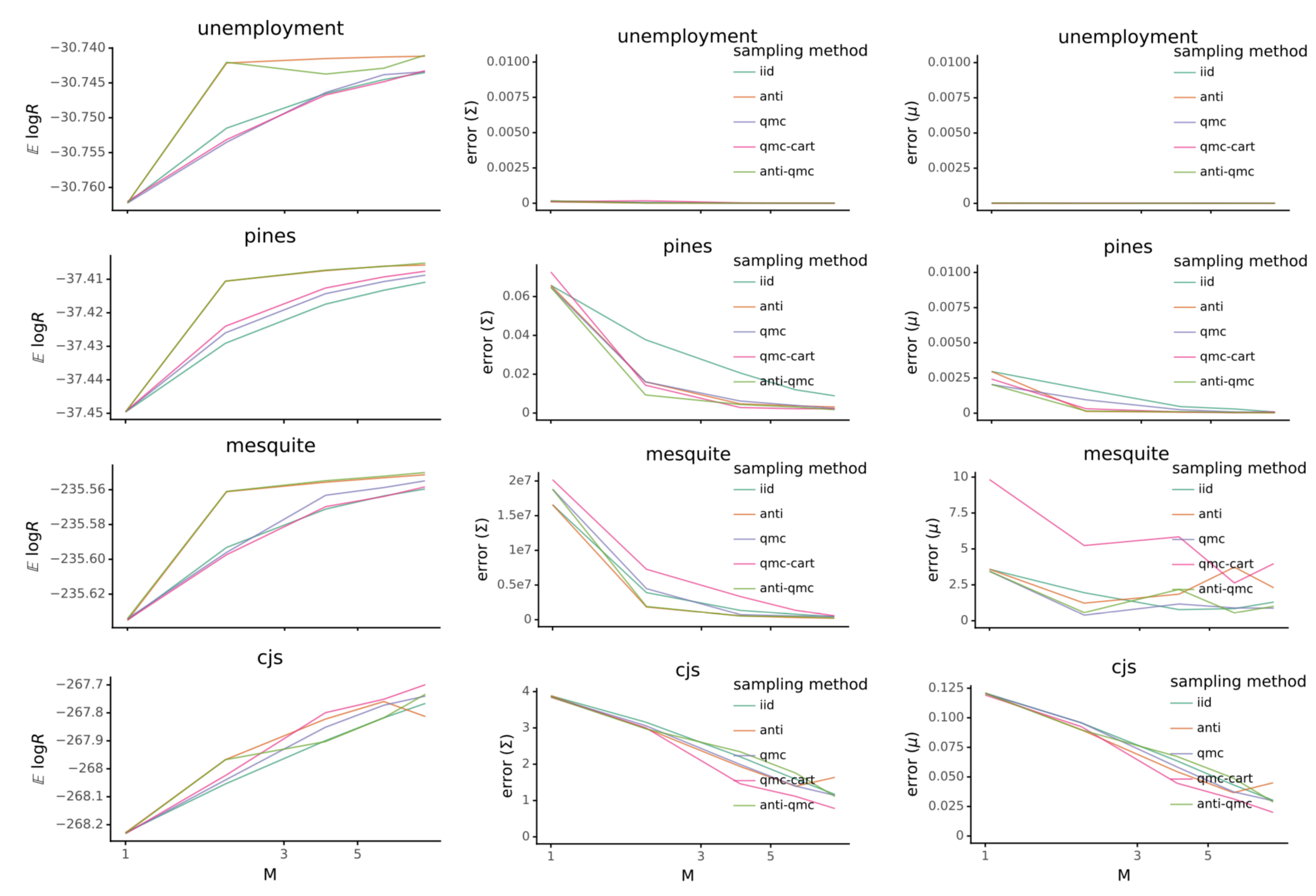
4. Example Implementation

Generate "batches" of samples from $[0,1]$ hypercubes and then transform.



5. Example Results

Other estimators may be more sample-efficient than iid (IWAEs)



Better likelihood bounds ⇔ better posteriors ✓