Adaptive Antithetic Sampling for Variance Reduction

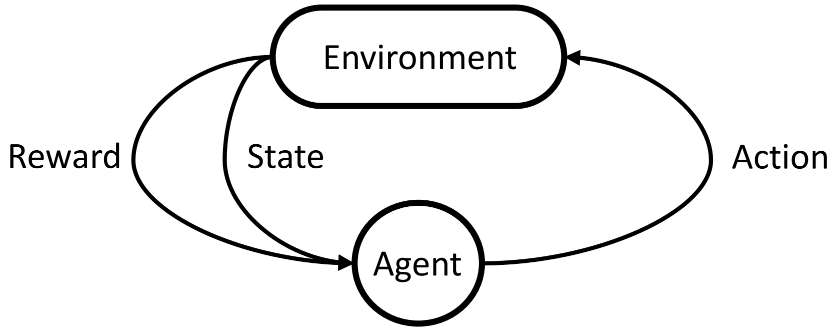
Stanford University

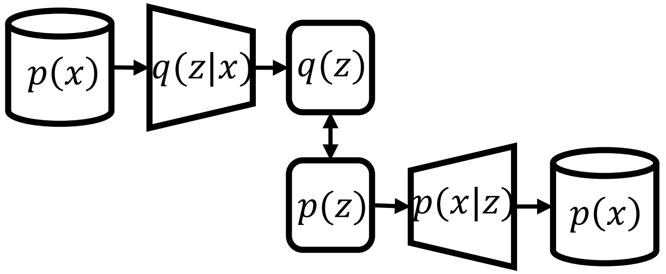
Hongyu Ren*, Shengjia Zhao*, Stefano Ermon

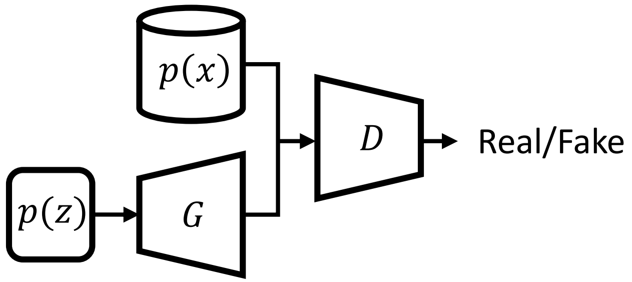
*equal contribution

Goal

Estimation of $\mu = \mathbb{E}_{p(x)}[f(x)]$ is ubiquitous in machine learning problems.







$$\mathbb{E}_{p(\tau)}\left[\sum_t r(s_t, a_t)\right]$$

Reinforcement Learning

$$\mathbb{E}_{p(x)}\mathbb{E}_{q(z|x)}\left[\log \frac{p(x,z)}{q(z|x)}\right]$$

Variational Autoencoder

$$\mathbb{E}_{p(x)}[\log D(x)] + \mathbb{E}_{p(z)}\left[\log\left(1 - D\left(G(z)\right)\right)\right]$$

Generative Adversarial Nets

Goal

Estimation of $\mu = \mathbb{E}_{p(x)}[f(x)]$ is ubiquitous in machine learning problems.

Monte Carlo Estimation: $\mu \approx \frac{1}{2}(f(x_1) + f(x_2))$

$x_1, x_2 \overset{\text{i.i.d.}}{\sim} p(x)$

MC is unbiased: $\mathbb{E}\left[\frac{1}{2}(f(x_1) + f(x_2))\right] = \mu$

High variance
Estimation can be far off with small sample size

Goal

Estimation of $\mu = \mathbb{E}_{p(x)}[f(x)]$ is ubiquitous in machine learning problems.

Monte Carlo Estimation: $\mu \approx \frac{1}{2}(f(x_1) + f(x_2))$

$x_1, x_2 \overset{\text{i.i.d.}}{\sim} p(x)$

Trivial solution:
use more samples!

Better solution:
better sampling strategy than i.i.d.

Antithetic Sampling

Don't sample i.i.d. $x_1, x_2 \sim p(x_1)p(x_2)$

Sample correlated distribution $x_1, x_2 \sim q(x_1, x_2)$

Unbiased if

$$q(x_1) = p(x_1)$$
$$q(x_2) = p(x_2)$$

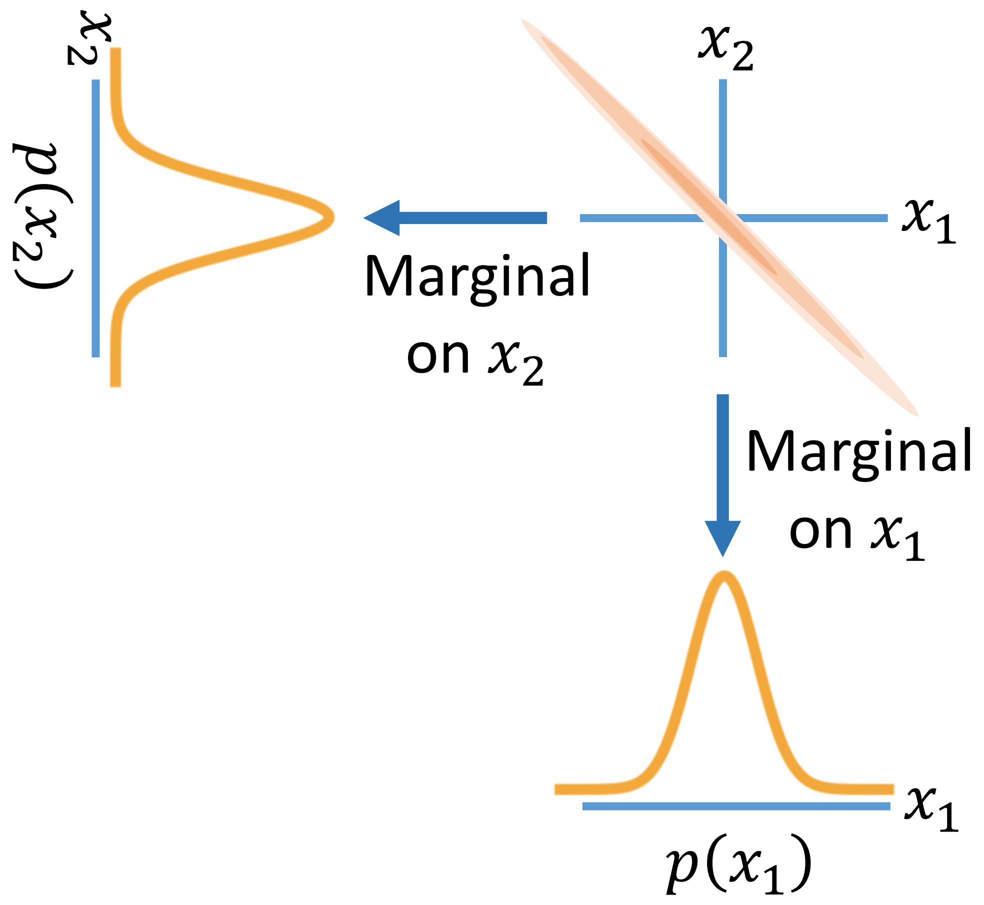
Goal: minimize

$$\mathrm{Var}_{q(x_1,x_2)}\left[\frac{f(x_1) + f(x_2)}{2}\right]$$

Example: Negative Sampling

$q(x_1, x_2)$ defined by

1. Sample $x_1 \sim p(x)$.
2. Pick $x_2 = -x_1$.

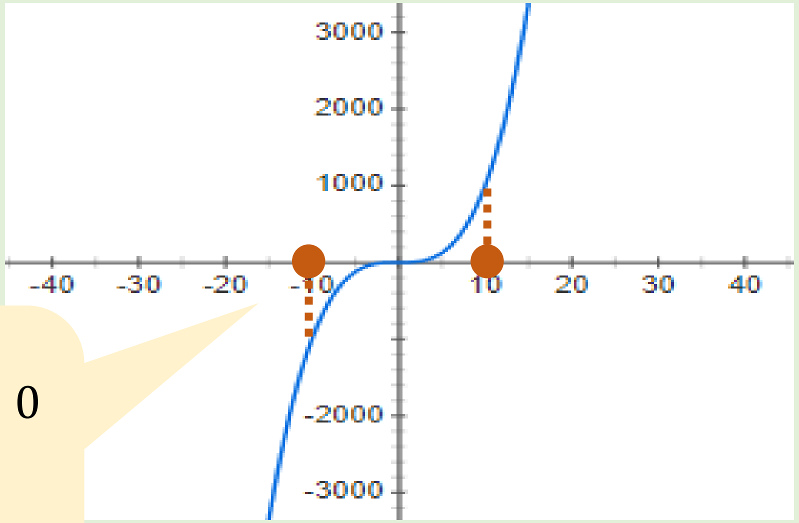


Example: Negative Sampling

$q(x_1, x_2)$ defined by

1. Sample $x_1 \sim p(x)$.
2. Pick $x_2 = -x_1$.

Best Case Example



$$\frac{f(x_1) + f(x_2)}{2} = 0$$

matches

$E_{p(x)}[f(x)] = 0$

$f = x^3$

$$\mathrm{Var}_{q(x_1,x_2)}\left[\frac{f(x_1)+f(x_2)}{2}\right] = 0$$

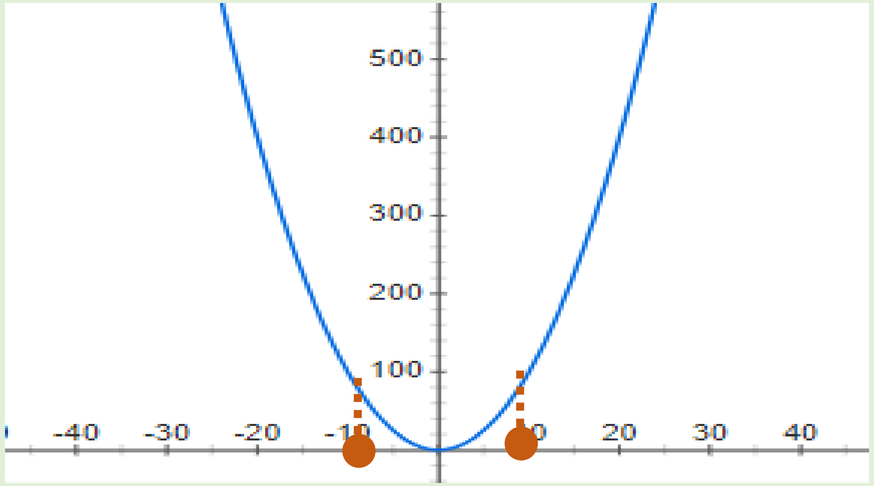
no error for a sample size of 2!

Example: Negative Sampling

$q(x_1, x_2)$ defined by

1. Sample $x_1 \sim p(x)$.
2. Pick $x_2 = -x_1$.

Worst Case Example



$f = x^2$

$f(x_1) = f(x_2), x_2$ redundant

$\mathrm{Var}_{q(x_1,x_2)}\left[\frac{f(x_1)+f(x_2)}{2}\right]$ doubles!

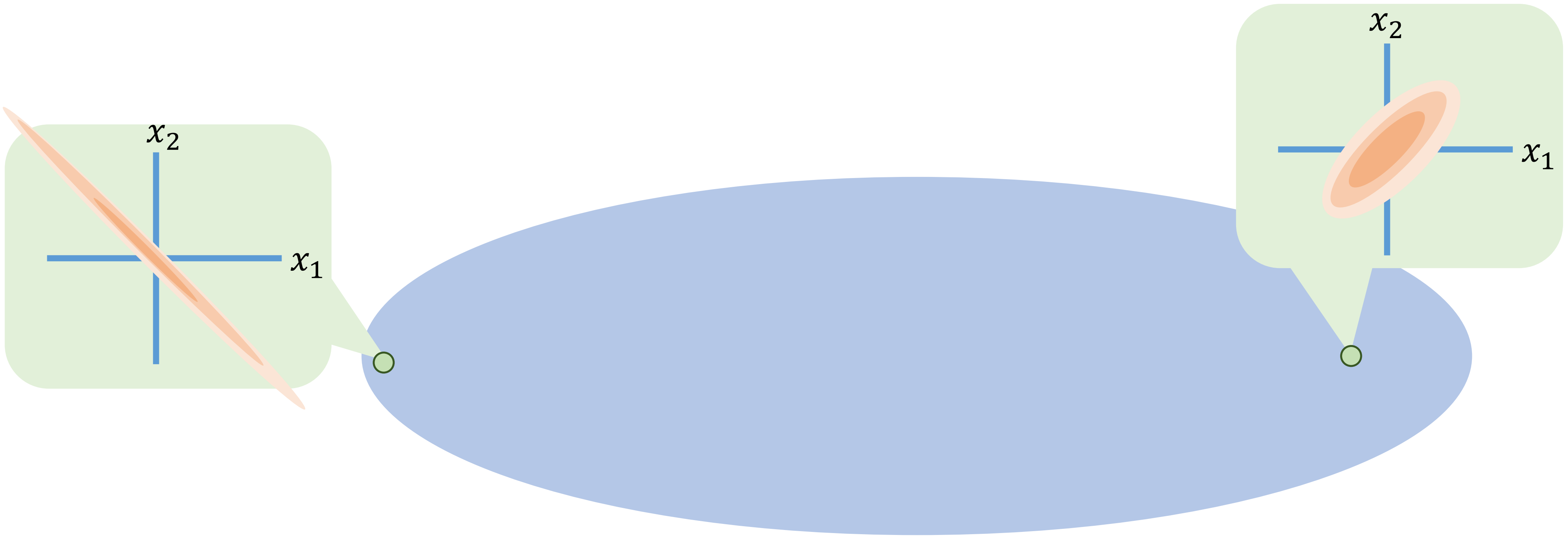
General Result

Question: is there an antithetic distribution that always works better than i.i.d.?

☺ Yes: sampling without replacement is always a tiny bit better.

☹ No Free Lunch (Theorem 1): no antithetic distribution work better than sampling without replacement for every function f.

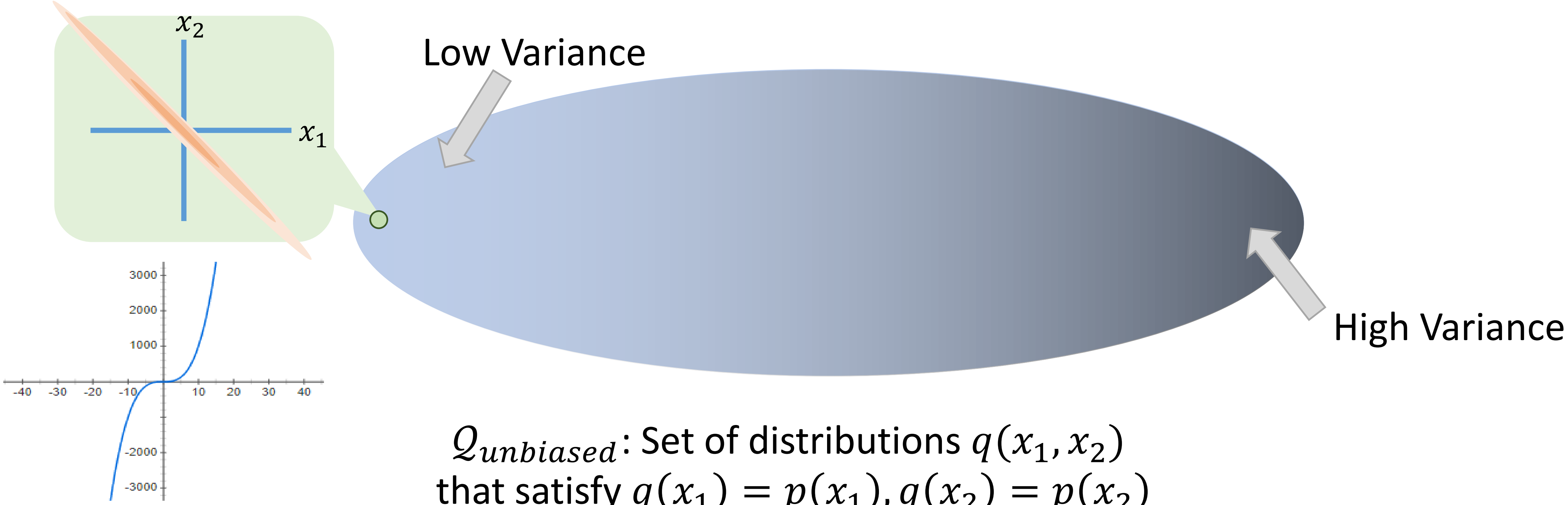
Valid Distribution Set



$\mathcal{Q}_{unbiased}$: Set of distributions $q(x_1, x_2)$
that satisfy $q(x_1) = p(x_1), q(x_2) = p(x_2)$

Variance of example functions

Pick this distribution

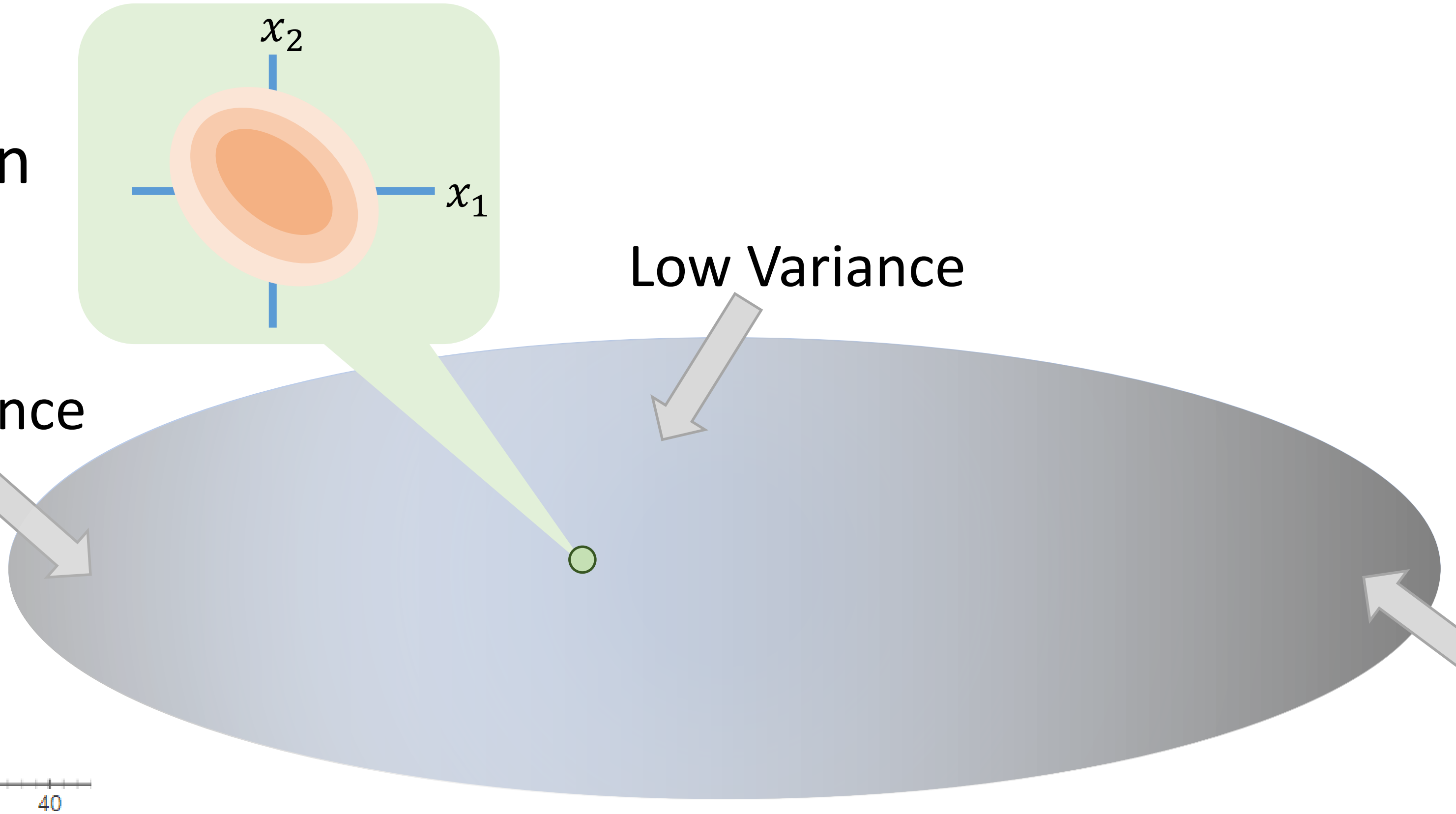


$f_1 = x^3$

$\mathcal{Q}_{unbiased}$: Set of distributions $q(x_1, x_2)$
that satisfy $q(x_1) = p(x_1), q(x_2) = p(x_2)$

Variance of example functions

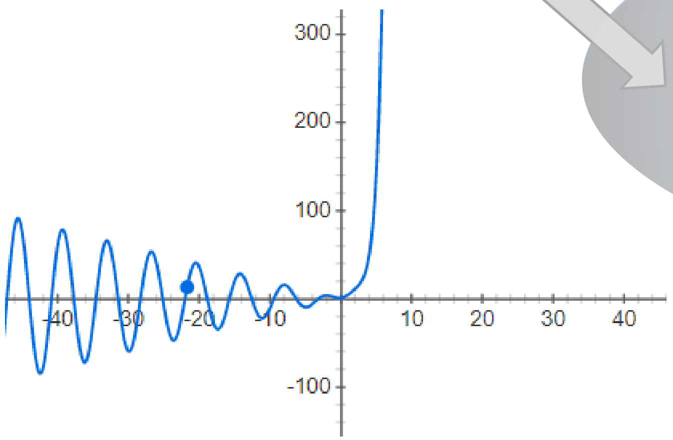
Pick this distribution



Low Variance

High Variance

High Variance

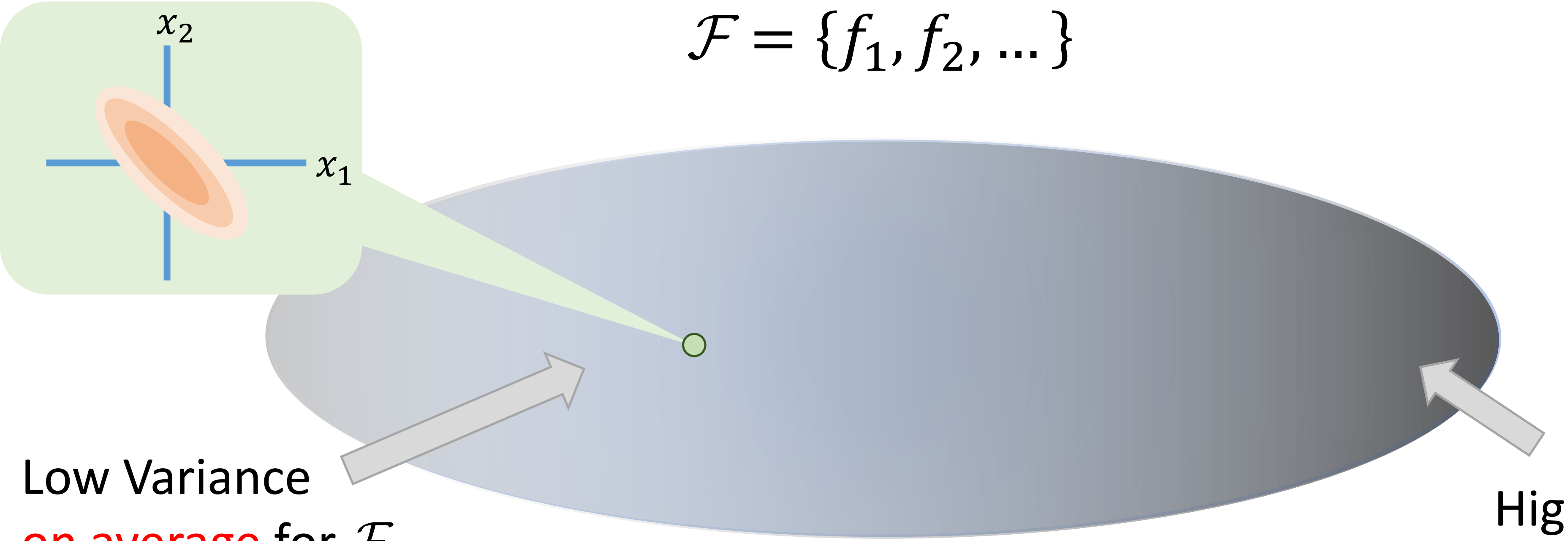


$f_2 = e^x + 2xsin(x)$

$\mathcal{Q}_{unbiased}$: Set of distributions $q(x_1, x_2)$
that satisfy $q(x_1) = p(x_1), q(x_2) = p(x_2)$

Pick Good Distribution for a Class of Functions

$\mathcal{F} = \{f_1, f_2, \ldots\}$

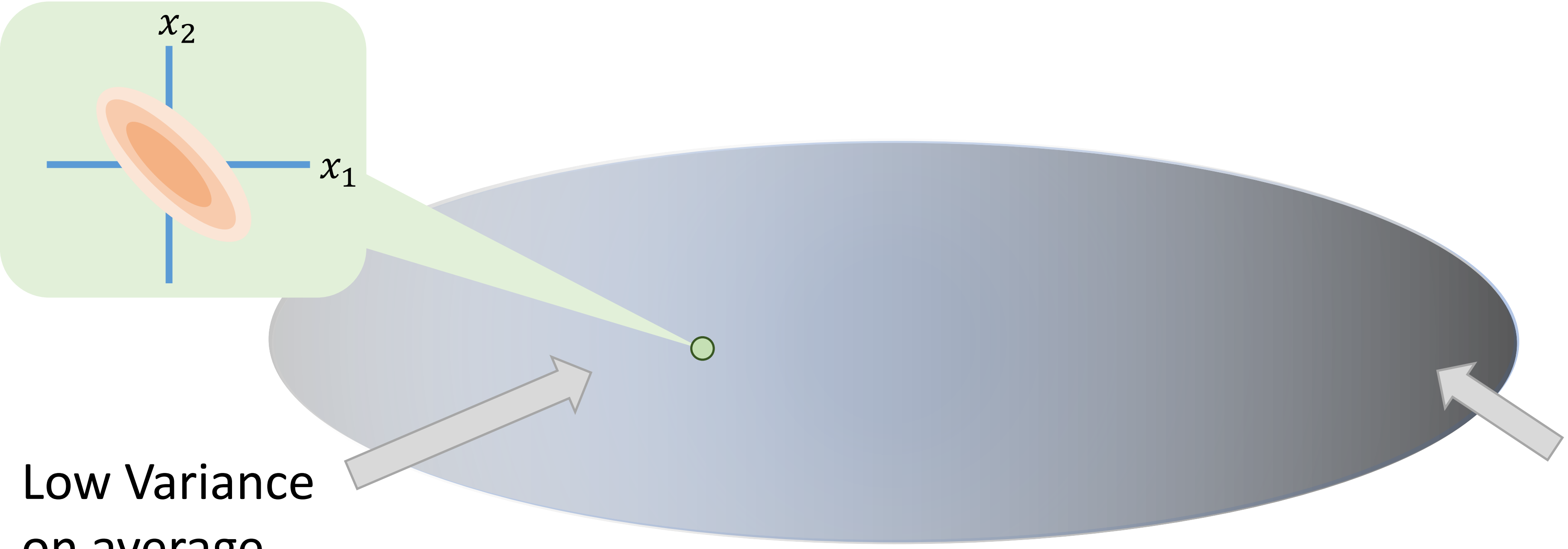


Low Variance
on average for $\mathcal{F}$

High Variance
on average for $\mathcal{F}$

$\mathcal{Q}_{unbiased}$: Set of distributions $q(x_1, x_2)$
that satisfy $q(x_1) = p(x_1), q(x_2) = p(x_2)$

Pick Good Distribution for a class of functions



$\mathcal{Q}_{unbiased}$: Set of distributions $q(x_1, x_2)$
that satisfy $q(x_1) = p(x_1), q(x_2) = p(x_2)$

Training
Pick a good q for several functions

Generalization
Low variance for similar functions

Training Objective

$$\min_{q} \mathbb{E}_{f\sim\mathcal{F}}\left[\mathrm{Var}_{q(x_1,x_2)}\left[\frac{f(x_1)+f(x_2)}{2}\right]\right]$$

$$s.t.\ q(x_1, x_2) \in \mathcal{Q}_{unbiased}$$

Practical Training Algorithm

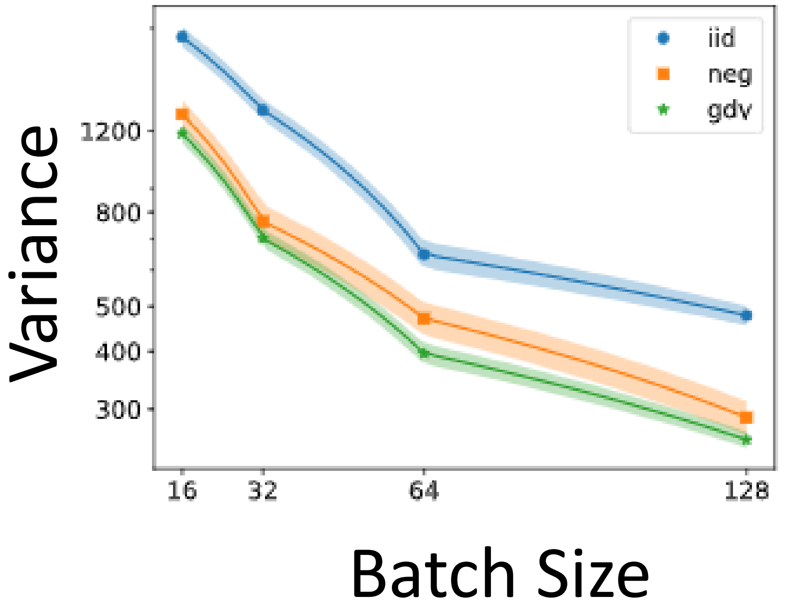
We design

1. Parameterization for $\mathcal{Q}_{unbiased}$ via copulas.

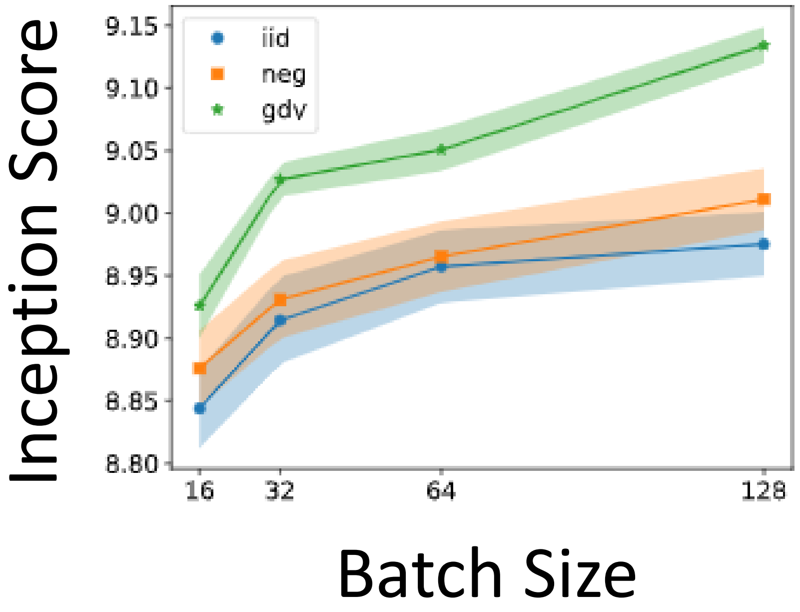
2. A surrogate objective to optimize the variance.

Wasserstein GAN w/ gradient penalty

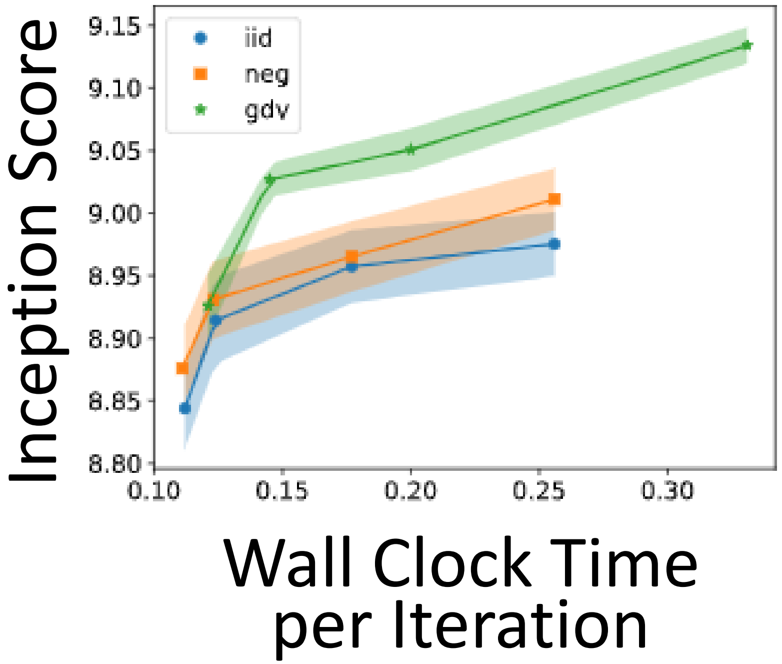
Variance of Gradient



Inception Score

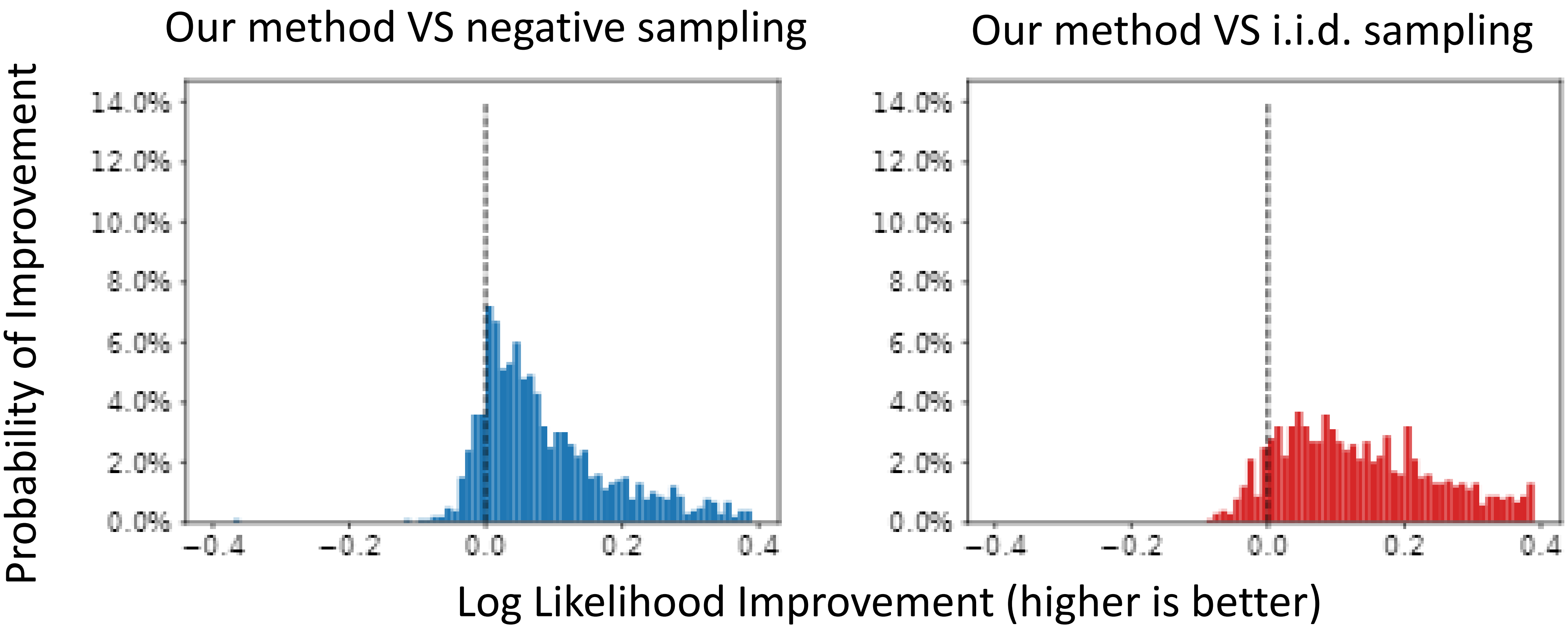


Inception Score



Gulrajani, Ishaan, et al. "Improved training of wasserstein gans." *Advances in Neural Information Processing Systems*. 2017.

Importance Weighted Autoencoder



Burda, Yuri, Roger Grosse, and Ruslan Salakhutdinov. "Importance weighted autoencoders." *arXiv preprint arXiv:1509.00519* (2015).

Conclusion

- Define a general family of (parameterized) unbiased antithetic distribution.
- Propose an optimization framework to learn the antithetic distribution based on the task at hand.
- Sampling from the resulting joint distribution reduces variance at negligible computation cost.

Welcome to our poster session for further discussions!

Thursday 6:30-9pm @ Pacific Ballroom #205