Variational Search Distributions

Daniel M. Steinberg, Rafael Oliveira, Cheng Soon Ong & Edwin V. Bonilla

Data61, CSIRO, Australia

{dan.steinberg, rafael.dossantosdeoliveira, cheng-soon.ong, edwin.bonilla}@data61.csiro.au

Openreview paper link

The Problem

We consider a variant of the active search problem [1], where we wish to find members (designs) of a rare desired class in a batch sequential manner with a fixed black-box evaluation (e.g. experiment) budget. We call sequential active learning of a *conditional generative* model of these designs **active generation**.

Examples of rare designs are compounds that could be useful pharmaceutical drugs, or highly active enzymes for catalyzing chemical reactions. We assume the design space is discrete or partially discrete, high-dimensional, and practically *innumerable*.

Contributions of this work:

1. formulate active generation objective for discrete design spaces,
2. present a variational inference algorithm, VSD, which solves this objective,
3. show that VSD performs well theoretically and empirically,
4. connect active generation to other generative methods for discrete BBO.

Formulating Active Generation

We are given a design space $\mathcal{X}$, which can be discrete or mixed discrete-continuous and high dimensional, and where for each instance that we choose $\mathbf{x} \in \mathcal{X}$, we measure some corresponding property of interest (so-called fitness) $y \in \mathbb{R}$,

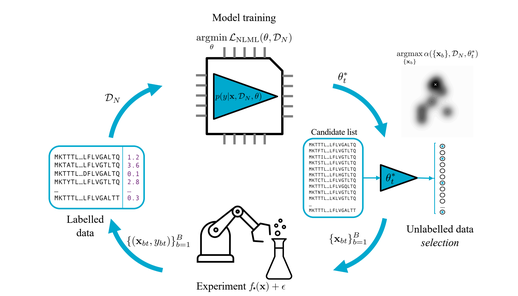
$$y = f_\bullet(\mathbf{x}) + \epsilon, \tag{1}$$

for some black-box function (e.g. the experiment), $f_\bullet$, and measurement error $\epsilon \in \mathbb{R}$, distributed according to $p(\epsilon)$ with $\mathbb{E}_{p(\epsilon)}[\epsilon] = 0$.

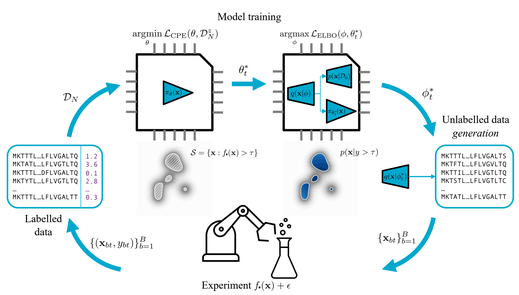
Instead of modeling the whole space, $\mathcal{X}$, we are only interested in a set of events based on fitness, $\mathcal{S} \subset \mathcal{X}$. We wish to learn a generative model, $q(\mathbf{x})$, that only returns samples $\mathbf{x}^{(s)} \in \mathcal{S}$, by efficiently querying the black-box function in Eqn. 1. For this work we consider primarily $\mathcal{S}_{\text{SLS}} := \{\mathbf{x} : f_\bullet(\mathbf{x}) > \tau\}$, but also $\mathcal{S}_{\text{BBO}} := \operatorname{argmax}_{\mathbf{x}} f_\bullet(\mathbf{x})$, though we only have access to noisy measurements, y.

Our primary task is active learning of the super level-set distribution, $p(\mathbf{x}|y > \tau)$.

How is Active Generation Different from Bayesian Optimization?



(a) *Unmodified* Batch BO, $\operatorname{argmin}_{\mathbf{x}} f_\bullet(\mathbf{x})$, select $\mathbf{x}$ from a list.



(b) VSD Active Generation, $\operatorname{argmin}_\phi \mathbb{D}[p(\mathbf{x}|y > \tau)\|q(\mathbf{x}|\phi)]$, generate $\mathbf{x}$.

Variational Search Distributions

We cast the estimation of $p(\mathbf{x}|y > \tau)$ as a sequential optimization problem. A suitable objective for a round, t, is to minimize a divergence,

$$\phi_t^* = \operatorname*{argmin}_\phi \mathbb{D}[p(\mathbf{x}|y > \tau)\|q(\mathbf{x}|\phi)]\,. \tag{2}$$

If we use the reverse Kullback Leibler divergence, we can cast this as maximizing the evidence lower bound (ELBO),

$$\mathcal{L}_{\text{ELBO}}(\phi) = \mathbb{E}_{q(\mathbf{x}|\phi)}[\log p(y > \tau|\mathbf{x})] - \mathbb{D}_{\text{KL}}[q(\mathbf{x}|\phi)\|p(\mathbf{x}|\mathcal{D}_0)]\,, \tag{3}$$

we assume access to a prior distribution over the space of designs, $p(\mathbf{x}|\mathcal{D}_0)$, that may be informed from the data at hand (or pre-trained).

We note the relationship between $\log p(y > \tau|\mathbf{x})$ and the probability of improvement (PI) acquisition function in Bayesian optimization (BO),

$$\log p(y > \tau|\mathbf{x}) \approx \log p(y > \tau|\mathbf{x}, \mathcal{D}_N) = \log \mathbb{E}_{p(y|\mathbf{x},\mathcal{D}_N)}[\mathbb{1}[y > \tau]] = \log \alpha_{PI}(\mathbf{x}, \mathcal{D}_N, \tau)\,. \tag{4}$$

Here $\mathbb{1} : \{\text{false}, \text{true}\} \to \{0, 1\}$ is the indicator function and $p(y|\mathbf{x}, \mathcal{D}_N)$ is typically estimated using the posterior predictive distribution of a Gaussian process (GP).

We plug in log PI in place of the expected log likelihood, and refer to this strategy as GP-PI. We rewrite the ELBO accordingly,

$$\mathcal{L}_{\text{ELBO}}(\phi, \tau, \mathcal{D}_N) = \mathbb{E}_{q(\mathbf{x}|\phi)}[\log \alpha_{PI}(\mathbf{x}, \mathcal{D}_N, \tau)] - \mathbb{D}_{\text{KL}}[q(\mathbf{x}|\phi)\|p(\mathbf{x}|\mathcal{D}_0)]\,. \tag{5}$$

The method that maximizes the objective in Equation 5 we call VSD.

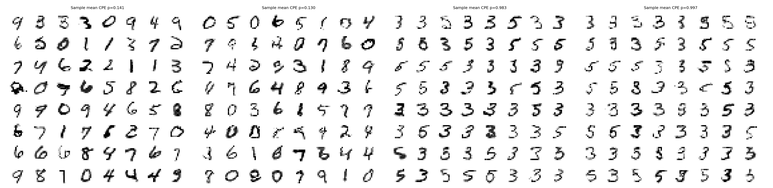
We use black-box variational inference [2] for optimization of our ELBO. This uses a Monte-Carlo *score function gradient estimator*, and allows us to avoid propagating gradients through discrete $\mathbf{x}$.

Class Probability Estimation

An alternative to GP-PI is to directly estimate the quantity $p(y > \tau|\mathbf{x}, \mathcal{D}_N)$. We do this with class probability estimator (CPE) on the labels $z := \mathbb{1}[y > \tau] \in \{0, 1\}$ so $p(y > \tau|\mathbf{x}, \mathcal{D}_N) = p(z = 1|\mathbf{x}, \mathcal{D}_N) \approx \pi_\theta(\mathbf{x})$, where $\pi_\theta : \mathcal{X} \to [0, 1]$. This is the same strategy used in BORE [3]. The VSD objective using CPE becomes,

$$\mathcal{L}_{\text{ELBO}}(\phi, \theta) = \mathbb{E}_{q(\mathbf{x}|\phi)}[\log \pi_\theta(\mathbf{x})] - \mathbb{D}_{\text{KL}}[q(\mathbf{x}|\phi)\|p(\mathbf{x}|\mathcal{D}_0)]\,, \tag{6}$$

and we use $\theta_t^* = \operatorname{argmin}_\theta \mathcal{L}_{\text{CPE}}(\theta, \mathcal{D}_N^z)$. CPE-PI allows VSD to scale $|\mathcal{D}_N| \gg \mathcal{O}(1000)$.



(a) LSTM Prior (b) Transformer Prior (c) LSTM Posterior (d) Transformer Post.

Figure: VSD conditioning an LSTM and transformer trained on MNIST to generate $\{3, 5\}$.

Covergence Rates

***Theorem* 2.1.** *Let assumptions D.1 to D.5 hold. Then VSD equipped with GP-PI approaches the level-set distribution at the following rate:*

$$\mathbb{D}[p(\mathbf{x}|y > \tau_t, \mathcal{D}_t)\|p(\mathbf{x}|y > \tau_t, f_\bullet)] \in \mathcal{O}_{\mathbb{P}}(t^{-1/2})\,. \tag{7}$$

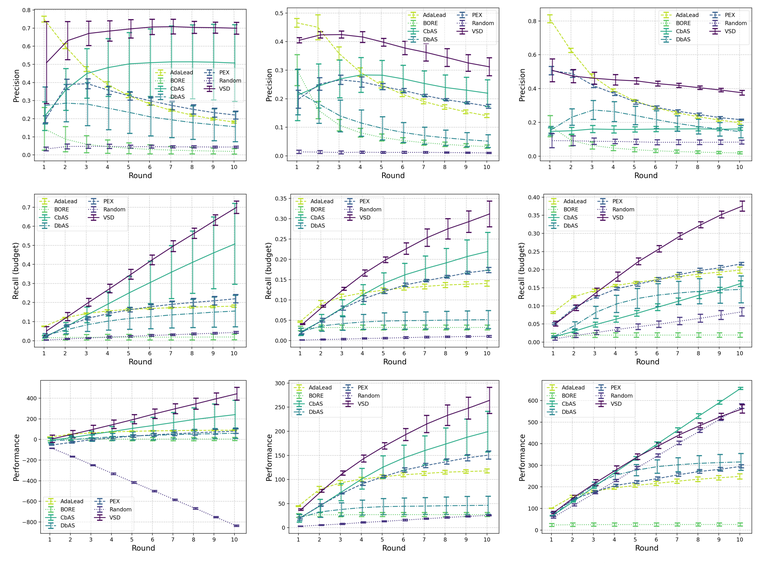
***Corollary* 2.2.** *Let π_θ be a fully connected ReLU network. Then, under assumptions in Proposition E.2, VSD with CPE-PI achieves:*

$$\mathbb{D}[p(\mathbf{x}|y > \tau_t, \mathcal{D}_t)\|p(\mathbf{x}|y > \tau, f_\bullet)] \in \widetilde{\mathcal{O}}_{\mathbb{P}}\left(t^{-\frac{1}{2(M+1)}}\right), \tag{8}$$

which asymptotically vanishes for all finite sequence lengths M.

Experiments

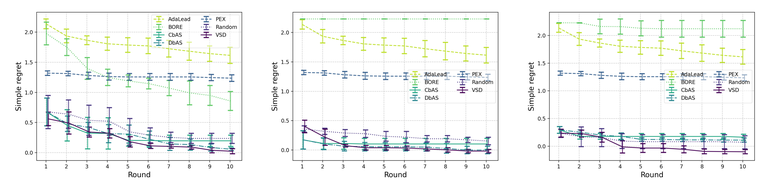
Fitness landscapes — fixed τ for learning $p(\mathbf{x}|y > \tau)$



(a) DHFR (b) TrpB (c) TFBIND8

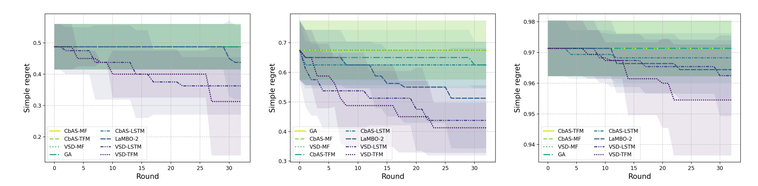
Figure: **Fitness landscape results**. Precision, recall and cumulative fitness (higher is better) on three combinatorially complete datasets. These measures are defined w.r.t. membership in the measured super level-set, $\mathbb{1}[y > \tau]$.

Black box optimization — increasing τ_t for finding $\operatorname{argmax}_{\mathbf{x}} f_\bullet(\mathbf{x})$



(a) Independent (b) LSTM (c) Transformer

Figure: **Black-box optimization results**. Simple regret (lower is better) on GFP with independent and auto-regressive variational distributions, $q(\mathbf{x}|\phi)$.



(a) $M = 15$ (b) $M = 32$ (c) $M = 64$

Figure: **Ehrlich function BBO results** (poli implementation). VSD and CbAS with different variational distributions; mean field (MF), LSTM and transformer (TFM), compared against genetic algorithm (GA) and LaMBO-2 baselines.

References

[1] Roman Garnett, Yamuna Krishnamurthy, Xuehan Xiong, Jeff G. Schneider, and Richard P. Mann. Bayesian optimal active search and surveying. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012.

[2] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial intelligence and statistics*, pages 814--822. PMLR, 2014.

[3] Louis C Tiao, Aaron Klein, Matthias W Seeger, Edwin V Bonilla, Cedric Archambeau, and Fabio Ramos. Bore: Bayesian optimization by density-ratio estimation. In *International Conference on Machine Learning*, pages 10289--10300. PMLR, 2021.