Stochastic Multiple Target Sampling Gradient Descent

Hoang Phan, Ngoc Tran, Trung Le, Toan Tran, Nhat Ho, Dinh Phung

Introduction

We propose *Stochastic Multiple Target Sampling Gradient Descent* (MT-SGD), which enables us to sample from multiple unnormalized target distributions. In summary, we make the following contributions in this work:

- Propose a principled framework that incorporates the power of Stein Variational Gradient Descent into multi-objective optimization. Concretely, we derive the formulation that extends the original work and allows sampling from multiple unnormalize distributions.
- Demonstrate our algorithm is readily applicable in the context of multi-task learning. The benefits of MT-SGD are twofold: i) the trained network is optimal, which could not be improved in any task without diminishing another, and ii) there is no need for predefined preference vectors as in previous works, MT-SGD implicitly learns diverse models universally optimizing for all tasks.
- Conduct comprehensive experiments to verify the behaviors of MT-SGD and demonstrate the superiority of MT-SGD to the baselines in a Bayesian setting, with higher ensemble performances and significantly lower calibration errors.

Multi-Target Sampling Gradient Descent

Given a set of target distributions $p_{1:K}(\theta) := \{p_1(\theta), \dots, p_K(\theta)\}$ with parameter $\theta \in \mathbb{R}^d$, we aim to find the optimal distribution $q^* \in \mathcal{Q}$ that minimizes a vector-valued objective function whose k-th component is $D_{KL}(q\|p_k)$:

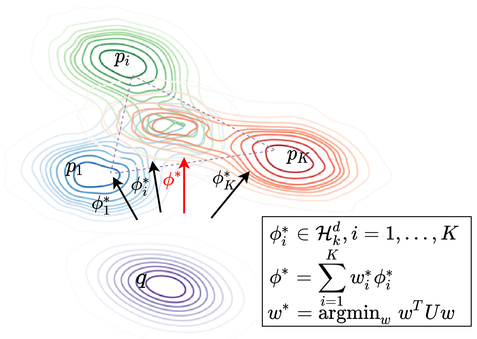
$$\min_{q \in \mathcal{Q}} \left[D_{KL}\left(q\|p_1\right), ..., D_{KL}\left(q\|p_K\right)\right], \tag{1}$$

Let us denote $\mathcal{H}_k$ by the Reproducing Kernel Hilbert Space (RKHS) associated with a positive semi-definite (p.s.d.) kernel k, and $\mathcal{H}_k^d$ by the d-dimensional vector function:

$$f = [f_1, \dots, f_d], (f_i \in \mathcal{H}_k).$$

Inspired by SVGD, at each step, assume that q is the current obtained distribution and the goal is to learn a transformation $T = id + \epsilon\phi$ so that $q^{[T]} = T\#q$ moves closer to $p_{1:K}$ simultaneously:

$$\min_{\phi} \left[D_{KL}\left(q^{[T]}\|p_1\right), ..., D_{KL}\left(q^{[T]}\|p_K\right)\right]. \tag{2}$$



For each target distribution p_i, the steepest descent direction is $\phi_i^* = \psi_i$, where $\psi_i(\cdot) = \mathbb{E}_{\theta\sim q}[k(\theta, \cdot)\nabla_\theta \log p_i(\theta) + \nabla_\theta k(\theta, \cdot)]$ and $\langle\cdot,\cdot\rangle_{\mathcal{H}_k^d}$ is the dot product in the RKHS. The KL divergence of interest $D_{KL}\left(q^{[T]}\|p_i\right)$ thus gets decreased roughly by $-\epsilon\left\|\phi_i^*\right\|^2_{\mathcal{H}_k^d}$ toward the target distribution p_i. The key steps of our MT-SGD are summarized in Algorithm 1

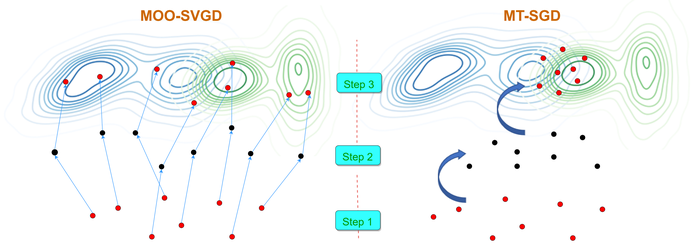
Algorithm 1 Pseudocode for MT-SGD.

Require: Multiple unnormalized target densities $p_{1:K}$.
Ensure: The optimal particles $\theta_1, \theta_2, \dots, \theta_M$.

1: Initialize a set of particles $\theta_1, \theta_2, \dots, \theta_M \sim q_0$.
2: **for** $t = 1$ to L **do**
3: Form the matrix $U \in \mathbb{R}^{K\times K}$.
4: Solve the QP $\min_{w\in\Delta_K} w^T U w$ to find the optimal weights $w^* \in \Delta_K$.
5: Compute the optimal direction $\phi^*(\cdot) = \sum_{i=1}^K w_i^* \phi_i^*(\cdot)$, .
6: Update $\theta_i = \theta_i + \epsilon\phi^*(\theta_i), i = 1, ..., K$.
7: **end for**
8: **return** $\theta_1, \theta_2, \dots, \theta_M$.

Related work

The most closely related work to ours is MOO-SVGD. In a nutshell, our MT-SGD navigates the particles from one distribution to another distribution consecutively with a theoretical guarantee of globally getting close to multiple target distributions.



By contrast, while MOO-SVGD also uses the MOO to update the particles, their employed repulsive term encourages the particle diversity without any theoretical-guaranteed principle to control the repulsive term, hence it can force the particles to scatter on the multiple distributions

Application to Multi-Task Learning

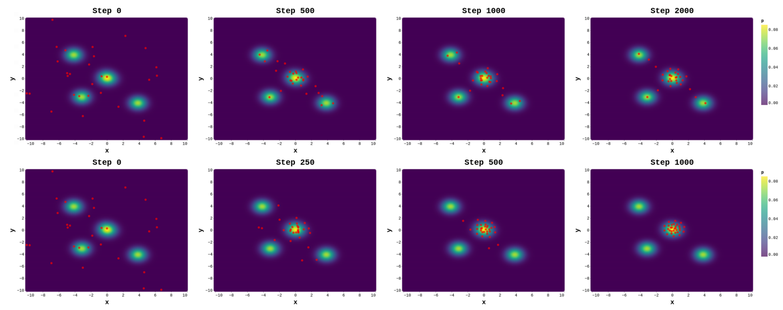
For multi-task learning, we assume to have K tasks $\{\mathcal{T}_i\}_{i=1}^K$ and a training set $\mathbb{D} = \{(x_i, y_{i1}, ..., y_{iK})\}_{i=1}^N$, where x_i is a data example and $y_{i1}, ..., y_{iK}$ are the labels for the tasks. The model for each task $\theta^j = \left[\alpha, \beta^j\right], j = 1, ..., K$ consists of the *shared part* α and *non-shared* part β^j targeting the task j. The posterior $p\left(\theta^j \mid \mathbb{D}\right)$ for each task reads

$$p\left(\theta^j \mid \mathbb{D}\right) \propto p\left(\mathbb{D} \mid \theta^j\right) p\left(\theta^j\right) \propto \prod_{i=1}^N p\left(y_{ij} \mid x_i, \theta^j\right)$$
$$\propto \prod_{i=1}^N \exp\left\{-\ell\left(y_{ij}, x_i; \theta^j\right)\right\} = \exp\left\{-\sum_{i=1}^N \ell\left(y_{ij}, x_i; \theta^j\right)\right\},$$

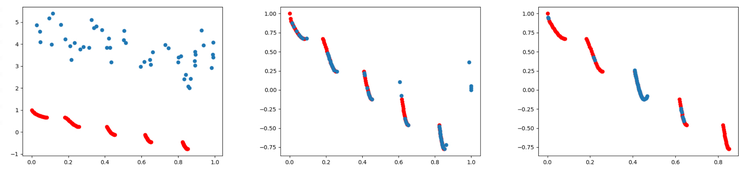
where ℓ is a loss function and the predictive likelihood $p(y_{ij} \mid x_i, \theta^j) \propto \exp\left\{-\ell\left(y_{ij}, x_i; \theta^j\right)\right\}$ is examined. Note that the prior $p\left(\theta^j\right)$ here is retained from previous studies, which is a uniform and non-informative prior and can be treated as a constant term in our formulation.

Experiments

We first qualitatively analyze the behavior of the proposed method on sampling from three target distributions. Each target distribution is a mixture of two Gaussians.



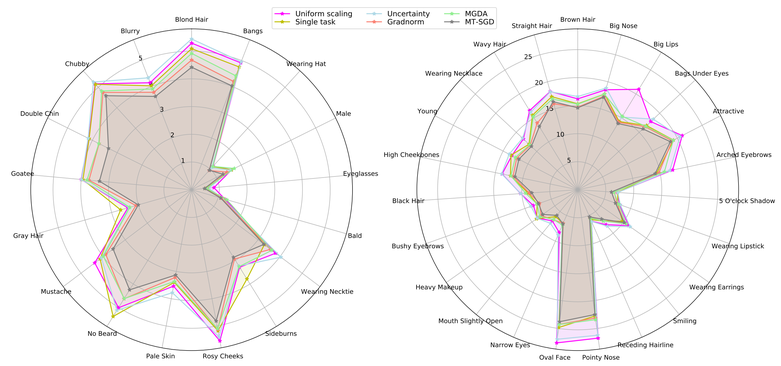
We next test our method on the other low-dimensional MOO problem. In particular, we use the two objectives ZDT3, whose non-contiguous Pareto front parts.



Our method is validated ondifferent benchmark datasets: (i) Multi-Fashion+MNIST, (ii) Multi-MNIST, and (iii) Multi-Fashion.

Dataset	Task	Linear scalarization	MGDA	Pareto MTL	MOO-SVGD	MT-SGD
Multi-Fashion+MNIST	Top left	21.33 ± 0.83	19.91 ± 0.26	9.44 ± 0.65	9.47 ± 0.89	**4.65 ± 0.11**
	Bottom right	17.76 ± 0.60	16.29 ± 1.35	4.73 ± 0.46	4.95 ± 0.49	**3.17 ± 0.20**
Multi-MNIST	Top left	17.37 ± 0.62	15.29 ± 0.49	5.45 ± 0.85	5.37 ± 0.51	**3.28 ± 0.20**
	Bottom right	18.09 ± 1.11	16.87 ± 0.67	7.34 ± 1.08	6.74 ± 0.50	**4.00 ± 0.19**
Multi-Fashion	Top left	15.86 ± 1.20	14.48 ± 0.95	8.55 ± 0.69	5.48 ± 0.53	**3.80 ± 0.38**
	Bottom right	15.98 ± 1.32	14.70 ± 1.63	9.01 ± 1.77	6.11 ± 0.54	**4.47 ± 0.21**

We performed our experiments on the CelebA dataset, which contains images annotated with 40 binary attributes.



As a final remark in the Multi-Fashion+Multi-MNIST experiment, we compare our methods against baselines in terms of the required running time