Policy Optimization as Wasserstein Gradient Flows

Ruiyi Zhang\textsuperscript{1}, Changyou Chen\textsuperscript{2}, Chunyan Li\textsuperscript{1}, and Lawrence Carin\textsuperscript{1}

\textsuperscript{1}Duke University  \hspace{1cm} \textsuperscript{2}SUNY at Buffalo

INTRODUCTION

Main Ideas:
- Formulate policy optimization as Wasserstein gradient flows.
- Stochastic policies optimized on the space of probability measures.
- Define WGFs for policy optimization in two settings.
  i) Indirect-policy learning: defined on a distribution space for parameters;
  ii) Direct-policy learning: defined on a distribution space for policy distributions.

Contributions:
- Place policy optimization into the space of probability measures, and interpret it as Wasserstein gradient flows (WGFs).
- Efficient algorithms by numerically solving the corresponding discrete gradient flows.
- A theoretically sound way to use (Wasserstein) trust-region.

BACKGROUND

Gradient flows on the probability-measure space: A Riemannian geometry $\mathcal{P}(\Omega)$ is characterized by the length between two elements (two distributions), defined by the 2nd-order Wasserstein distance:

$$ W_2^2(\mu, \nu) \triangleq \inf_{\{\gamma \in \Gamma(\mu, \nu)\}} \mathbb{E}_\gamma \|x - y\|^2 \mathbb{P}(x, y) : \gamma \in \Gamma(\mu, \nu), $$

where $\Gamma(\mu, \nu)$ is the set of all transport maps between $\mu$ and $\nu$, and $\mathbb{E}_\gamma$ denotes the expectation with respect to $\gamma$.

Function $F$ mapping a probability measure $\mu$ to a real value, i.e., $F : \mathcal{P}(\Omega) \to \mathbb{R}$, with

$$ F(\mu) \triangleq \mathbb{E}_\mu \ln \frac{\phi}{\mathbb{P}} + \mathbb{E}_\mu \phi, $$

where $\mathbb{P}$ is the target distribution. The energy functional $F$ of an Itô diffusion is defined as

$$ F(\mu) \triangleq -\int \mathbb{P}(x) \ln \mathbb{P}(x) dx + \int \mathbb{P}(x) \phi(x) dx. $$

Numerical Solution: WGFs can be approximately solved by discrete-time flows, formulated as an iterative optimization problem:

$$ \mu^{(h)}_k = \arg \min_{\mu \in \mathcal{P}(\Omega)} F(\mu), $$

where $\mu^{(h)}_0$ is the target distribution. This procedure is called the Jordan-Kinderlehrer-Otto (JKO) scheme.

Challenge and Solution:
- Original WGFs is infeasible to directly optimize over a distribution.
- Particle-based methods to approximate a continuous density function.

EXPERIMENTS

Indirect Policy Learning

Figure: Learning curves by IP-WGF and SVG with REINFORCE and A2C.

Direct Policy Learning

Figure: Average return in MuJoCo tasks by Soft-Q, SAC and DP-WGF-V (first row), and by DDPG, TRPO-GAE and DP-WGF-V (second row).

<table>
<thead>
<tr>
<th>Domain</th>
<th>Threshold</th>
<th>Episodic Return (Episode)</th>
<th>Episodic Return (Episode)</th>
<th>Episodic Return (Episode)</th>
<th>Episodic Return (Episode)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walker</td>
<td>300</td>
<td>4978.59</td>
<td>2289</td>
<td>4255.05</td>
<td>388</td>
</tr>
<tr>
<td>Hopper</td>
<td>3000</td>
<td>3248.76</td>
<td>678</td>
<td>3146.51</td>
<td>738</td>
</tr>
<tr>
<td>Humanoid</td>
<td>2000</td>
<td>3077.84</td>
<td>18740</td>
<td>2212.51</td>
<td>2465</td>
</tr>
</tbody>
</table>

Table: WGF-DP-V, SAC, and DDPG results showing the max average rewards attained and the episodes to cross specific reward thresholds. WGF-DP-V often learn more sample-efficiently than the baselines, and WGF-DP-V can solve difficult domains such as Humanoid better than DDPG.