

輸送写像による確率測度の最適化とその応用

東京大学 二反田 篤史

東京大学 鈴木 大慈

機械学習では確率測度についての最適化問題がしばしば現れ、その汎用解法の開発は重要である。本研究では、有限次元空間における勾配法あるいは確率的勾配法を拡張する形で確率測度の最適化法を提案し、その収束性を示す。また、応用例として確率測度による分類器のアンサンブル学習と生成モデル学習を取り上げ、提案手法の適用方法とその際の利点を簡単に説明する。

確率測度の最適化

最適化対象の変数はボレル可測空間 $\Theta = \mathbb{R}^d$ 上の確率測度 μ であり、目的関数を \mathcal{L} とすれば、本研究が対象とする最適化問題は $\min_{\mu} \mathcal{L}(\mu)$ で表される。この問題に対する提案手法の反復では、確率測度 μ は輸送写像 $\psi : \text{supp}(\mu) \rightarrow \mathbb{R}^d$ を用いて $\psi_{\#}\mu$ に更新される。ここで、 $\psi_{\#}\mu$ は確率測度の押し出しである。特に ψ として $\text{supp}(\mu)$ 上の滑らかなベクトル場 $\xi : \text{supp}(\mu) \rightarrow \mathbb{R}^d$ による摂動 $\psi = id + \xi$ を考える。この時、得られる確率測度は $\mu_k = (id + \xi_{k-1})_{\#}\mu_{k-1} = ((id + \xi_{k-1}) \circ \dots \circ (id + \xi_0))_{\#}\mu_0$ という形をとり、最適化手法を構築するにあたり考えるべきは、各反復においての摂動 ξ_j の選び方となる。初期確率測度 μ_0 としてディラック測度 δ_{θ_0} を採用した場合、この方法は有限次元ユークリッド空間での最適化問題と等価になる。実際、ディラック測度 δ_{θ} とユークリッド空間における点 $\theta \in \Theta$ の同一視を用いれば、目的関数は θ に依存する関数 $\mathcal{L}(\theta)$ であり、確率測度の反復: $\delta_{\theta_k} \leftarrow (id + \xi_{k-1})_{\#}\delta_{\theta_{k-1}} = \delta_{\theta_{k-1} + \xi_{k-1}(\theta_{k-1})}$ は $\theta_k \leftarrow \theta_{k-1} + \xi_{k-1}(\theta_{k-1})$ となる。従って、本研究の問題設定および提案手法は有限次元ユークリッド空間での最適化を特殊な場合として含む。そして、その場合に知られる有用な知見のいくつかを我々の問題へ一般化する事で、具体的な最適化手法の構築及び理論解析の手がかりとする事が出来る。特に勾配法に対応するものとして、摂動を $\mathcal{L}((id + \xi)_{\#}\mu)$ の ξ についてのフレシェ微分 $\nabla_{\xi} \mathcal{L}((id + \xi)_{\#}\mu)|_{\xi=0}$ とする事は自然である。この場合、適当な仮定の下で $L^2(\mu)$ -内積による以下の等式が成立する。

$$\mathcal{L}((id + \xi)_{\#}\mu) = \mathcal{L}(\mu) + \langle \nabla_{\zeta} \mathcal{L}((id + \zeta)_{\#}\mu)|_{\zeta=0}, \xi \rangle_{L^2(\mu)} + O(\|\xi\|_{L^2(\mu)}^2).$$

これは、摂動についてのテイラーの公式に他ならず、有限次元の性質を一般化する上で重要な等式である。実際、 $\xi = -\nabla_{\zeta} \mathcal{L}((id + \zeta)_{\#}\mu)|_{\zeta=0}$ が μ における降下方向である事が直ちに分かる。従って、フレシェ微分あるいは、その推定量を用いた降下法により確率測度についての最適化が実行される。また、本研究ではフレシェ微分のノルムが確率測度の集合上で最適性ギャップとして適切な性質を備えている事を示し、本手法が下記の局所最適性条件を満たす解へ収束する事を紹介する。

$$\|\nabla_{\xi} \mathcal{L}((id + \xi)_{\#}\mu)|_{\xi=0}\|_{L^2(\mu)}^2 = 0.$$

また、この条件の更なる正当化として、 $\mathcal{L}(\mu)$ の μ についての通常の意味での凸性とフレシェ微分の滑らかさの仮定の下、局所最適性条件を満たす確率測度はその台の内側でのみ体積を持つ確率測度の中で最適であるという性質を紹介する。

確率測度による分類器のアンサンブル学習

本手法の分類器のアンサンブル学習への適用法と、その性質を簡単に紹介する。 $\mathcal{X} \subset \mathbb{R}^m, \mathcal{Y} = \{-1, 1\}$ をデータが属す空間と二値ラベル集合とし、 $\{h(\theta, \cdot) : \mathcal{X} \rightarrow [-1, 1] \mid \theta \in \Theta\}$ を滑らかにパラメトライズされた弱分類器の集合とする。与えられた訓練データ $S = \{x_i, y_i\}_{i=1}^N$ から弱分類器集合上の確率測度 μ によるアンサンブル $h_{\mu}(x) = \mathbb{E}_{\mu}[h(\theta, x)]$ を得る事が目標である。この目標は損失関数 l が定める経験損失 $\mathcal{L}(\mu) = \mathbb{E}_{(x,y) \sim S}[l(-yh_{\mu}(x))]$ の最小化で達成され、そのために提案手法が適用可能である。本発表では、この方法が AdaBoost と違い早期終了を必要としない事を簡単に述べ、実験結果を紹介する。