機械学習における最適化理論と学習理論的側面
演習問題 (COSS2020)
2020 年 8 月 6 日

鈴木大慈
e-mail: taiji@mist.i.u-tokyo.ac.jp

問 1　必ずしも滑らかとは限らない凸関数を (劣) 勾配法で最小化するとき，定数ステップサイズを用いてしまうと収束しない例を作れ．

Definition 1 ある凸関数 $f:\mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$ の凸共役関数 (Legendre 変換) は $f^*(y) = \sup_{x\in\mathbb{R}^d}\{x^\top y - f(x)\}$ $(y \in \mathbb{R}^d)$ で定義される．

問 2　$f(x) = f^*(x)$ $(\forall x \in \mathbb{R}^d)$ となる凸関数を求めよ．これは一意に定まるか？

問 3　$V = \{1,\dots,d\}$ とする．非負集合関数 $F: 2^V \to \mathbb{R}_+ \cup \{\infty\}$[*1]で，$F(\emptyset) = 0$, $F(A) > 0$ $(\forall A \subset V,\ A \neq \emptyset)$ を満たすものを考える．$p \geq 1$ として，$q \geq 1$ を $\frac{1}{p} + \frac{1}{q} = 1$ を満たすものとする．インデックス集合 $A \subset V$ に対して，$s \in \mathbb{R}^d$ の A への制限 $s_A \in \mathbb{R}^d$ を $s_{A,i} = 0$ $(i \notin A)$, $s_{A,i} = s_i$ $(i \in A)$ とする．また，$s(A) := \sum_{i\in A} s_i$ とする．集合関数 F に対して，$\mathbb{R}^d$ 上のノルム Ω_p とその双対ノルム $\Omega_p^\bullet$ を以下のように定める [3]:

$$\Omega_p^\bullet(s) := \max_{A\subset V, A\neq\emptyset} \frac{\|s_A\|_q}{F(A)^{1/q}}\ (s \in \mathbb{R}^d),$$
$$\Omega_p(w) := \sup\{w^\top s \mid s \in \mathbb{R}^d,\ \Omega_p^\bullet(s) \leq 1\}\ (w \in \mathbb{R}^d).$$

(i) $\Theta(w) := F(\mathrm{supp}(w))^{1/q}\|w\|_p$ としたとき，Ω_p は Θ の凸包であることを示せ．つまり，$\Theta^{**} = \Omega_p$ であることを示せ．

(ii) $F(A) = |A|$ に対して，Ω_p を求めよ．

(iii) $F_-(A) = \max\{s(A) \mid s \in \mathbb{R}_+^d,\ s(B) \leq F(B)\ (\forall B \subset V)\}$ とする．$\Omega_p^F = \Omega_p^{F_-}$ を示せ (なお，Ω_p^F は集合関数 F に対応するノルム Ω_p である)．

(iv) $\Omega_p(w) = \min_{\eta\in\mathbb{R}_+^d} \sum_{i\in V} \frac{1}{p}\frac{|w_i|^p}{\eta_i^{p-1}} + \frac{1}{q}\Omega_\infty(\eta)$ であることを示せ (ヒント: $|a|^{1/q}|w| = \min_{\eta\in\mathbb{R}_+} \frac{1}{p}\frac{|w|^p}{\eta^{p-1}} + \frac{1}{q}\eta|a|$ $(a \in \mathbb{R},\ w \in \mathbb{R})$)．これより，近接写像が

$$\min_{w\in\mathbb{R}^d} \frac{1}{2}\|x - w\|^2 + \lambda\Omega_2(w) = \frac{\lambda}{2}\min_{\eta\in\mathbb{R}_+^d}\left(\sum_{i\in V}\frac{x_i^2}{\eta_i + \lambda} + \Omega_\infty(\eta)\right)$$

で与えられることを示せ．

[*1] $\mathbb{R}_+ := \{x \in \mathbb{R} \mid x \geq 0\}$

(v) $d_i > 0$ $(i \in V)$ に対して，$F(A) = \sum_{i=1}^{\max(A)} d_i$ とする．このとき，

$$\Omega_2(w) = \min_{\eta \in \mathbb{R}_+^d} \left\{ \frac{1}{2} \sum_{i \in V} \left[\frac{w_i^2}{\eta_i} + \eta_i d_i \right] \ \middle| \ \eta_i \le \eta_{i-1} \ (\forall i) \right\}$$

を示せ (wedge penalty と呼ばれる)．

(vi) F が単調劣モジュラ関数のとき，その Lovász 拡張と Ω_∞ との関係を考察せよ．

問 4 $a_1, \ldots, a_n > 0$, $\alpha, \beta > 0$ とする．次の最適化問題の解を求めよ:

$$\min_x \left\{ \sum_{i=1}^n \frac{a_i}{x_i^\alpha} \ \middle| \ x_i > 0,\ \sum_{i=1}^n x_i^\beta \le 1 \right\}.$$

問 5 γ-平滑でかつ μ-強凸関数を Nesterov の加速法のリスタート法を用いて最適化する．このとき，$f(x_t) - f(x^*) = \mathrm{O}(\exp(-t\sqrt{\mu/\gamma}))$ を示せ．なお，講義中で紹介した γ-平滑な凸関数における加速法の収束レートは用いて良い．

Theorem 2 (Fenchel の双対定理 (特殊ケース)) $f : \mathbb{R}^n \to \mathbb{R}$ と $g : \mathbb{R}^m \to \mathbb{R}$ を閉凸関数とする．$A \in \mathbb{R}^{n \times m}$ とする．このとき，以下が成り立つ:

$$\inf_{x \in \mathbb{R}^m} \{f(Ax) + g(x)\} = -\inf_{y \in \mathbb{R}^n} \{f^*(y) + g^*(-A^\top y)\}.$$

問 6 $(x_i, y_i)_{i=1}^n \in \mathbb{R}^n \times \{\pm 1\}$ なる訓練データ (二値判別問題 $(y_i \in \{\pm 1\})$) が与えられているとして，以下の Support Vector Machine (SVM) の学習問題を考える:

$$\inf_{\alpha \in \mathbb{R}^n} \ \frac{1}{n} \sum_{i=1}^n \max\left\{0, 1 - \sum_{j=1}^n \alpha_j k(x_i, x_j) y_i \right\} + \frac{\lambda}{2} \sum_{i=1}^n \sum_{j=1}^n k(x_i, x_j) \alpha_i \alpha_j.$$

ただし，$k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ はカーネル関数とする．Fenchel の双対定理より，この問題の双対問題を導出せよ．(ここでは簡単のためバイアス項を抜かしているが，余裕があればバイアス項有りの双対問題も導出されたい．つまり，$(\sum_j \alpha_j k(x_i, x_j)) y_i$ の項を $(\sum_j \alpha_j k(x_i, x_j) + b) y_i$ に修正して，$b \in \mathbb{R}$ についても最適化する問題を考える)

Theorem 3 (Knott-Smith optimality criterion, Brenier's theorem) W_2 -距離は以下のように定義される:

$$W_2(\mu, \nu) := \sqrt{\inf_{\pi \in \Pi(\mu, \nu)} \int (x - y)^2 \mathrm{d}\pi(x, y)}. \tag{1}$$

ここで，μ, ν は $\mathbb{R}^d$ 上の (Borel) 確率測度でともに二次モーメントが有界であるとし，Π は周辺分布が μ および ν である $\mathbb{R}^d \times \mathbb{R}^d$ 上の確率測度全体の集合である．今，μ, ν がルベーグ測度に関して絶対連続な確率測度であるなら，式 (1) の inf を達成する π が存在し，π が最適解であることの必要十分条件は，ある真閉凸関数 φ が存在して，π-a.s. で $y \in \partial\varphi(x)$ が成

り立つことである．すなわち，最適解は $\pi(x,y) = (\mathrm{Id}\times\nabla\varphi)_\sharp\mu$ で与えられる[*2]．なお，この写像 $T=\nabla\varphi$ は以下の最適輸送問題の μ-a.s. で一意な解である:

$$\inf_{T:T_\sharp\mu=\nu}\int_{\mathbb{R}^d}\|x-T(x)\|^2\mathrm{d}\mu(x) = W_2(\mu,\nu)^2.$$

（詳細は [5] を参照されたい）．

問 7 $\mu_1,\mu_2\in\mathbb{R}^d$ かつ，$\Sigma_1,\Sigma_2\in\mathbb{R}^{d\times d}$ をともに正定値対象行列とする．二つの正規分布 $N(\mu_1,\Sigma_1), N(\mu_2,\Sigma_2)$ の間の W_2-距離を求めよ． (直接導出することも可能 [1, 4])．

Remark 4 $(\Omega,\mathcal{F})$ を確率空間，$(\mathcal{F}_t)_{t\ge0}$ をその上のをフィルトレーションとする．確率微分方程式

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}W_t$$

を考える．ただし，$W=\{W_t\}_t$ は $\mathcal{F}_t$-適合な $\mathbb{R}^d$-値ブラウン運動で，$b:\mathbb{R}^d\to\mathbb{R}^d$, $\sigma:\mathbb{R}^d\to\mathbb{R}^{d\times d}$ はともに (十分滑らかな) ボレル可測関数である．この確率微分方程式に対し，無限小生成作用素 $\mathcal{A}$ は

$$\lim_{t\searrow0}\frac{1}{t}[\mathrm{E}[f(X_t)|X_0=x]-f(x)] = (\mathcal{A}f)(x) \quad (x\in\mathbb{R}^d)$$

で与えられる $(f\in C^2(\mathbb{R}^d))$．実は，$A(x)=\sigma(x)^\top\sigma(x)$ を用いて，$\mathcal{A}f(x) = \frac{1}{2}\sum_{i,j=1}^d A_{ij}(x)\frac{\partial^2 f(x)}{\partial x_i\partial x_j} + \sum_{i=1}^d b_i(x)\frac{\partial f(x)}{\partial x_i}$ と表せることが知られている．このとき，状態遷移確率 $P(X_t\in\mathrm{d}y|X_0=x) = \Gamma(t;xy)\mathrm{d}y$ に対し，前進 Kolmogorov の方程式 (前進 Fokker-Planck 方程式)

$$\frac{\partial}{\partial t}\Gamma(t;x,y) = \mathcal{A}^*\Gamma(t;x,y) \quad ((t,y)\in(0,\infty)\times\mathbb{R}^d),$$

および，後退 Kolmogorov の方程式 (後退 Fokker-Planck 方程式)

$$\frac{\partial}{\partial t}\Gamma(t;x,y) = \mathcal{A}\Gamma(t;x,y) \quad ((t,x)\in(0,\infty)\times\mathbb{R}^d),$$

が満たされることが知られている [2]．なお，$\mathcal{A}^*f(y) = \frac{1}{2}\sum_{i,j=1}^d\frac{\partial^2}{\partial y_i\partial y_j}[A_{ij}(y)f(y)] - \sum_{i=1}^d\frac{\partial}{\partial y_i}[b_i(y)f(y)]$ である．
（参考）伊藤の公式：$f:\mathbb{R}^d\to\mathbb{R}$ が二回微分可能なとき，

$$\mathrm{d}f(X_t) = \nabla^\top f(X_t)(b(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}B_t) + \frac{1}{2}\sum_{i,j=1}^d\frac{\partial^2 f(X_t)}{\partial x_i\partial x_j}A_{i,j}(X_t)\mathrm{d}t.$$

問 8 Kolmogorov の方程式から連続時間勾配ランジュバン動力学の不変測度 (定常分布) π_∞ を求めよ．ただし，その存在および一意性，不変測度への収束は仮定してよい．

[*2] $\nabla\varphi$ は必ずしも存在しないが，μ-a.s. で well-defined であることが知られている [5]．

問 9　連続時間勾配ランジュバン動力学の不変測度 π_∞ が対数 Sobolev 不等式を満たすとき，X_t の分布 ρ_t と π_t との相対エントロピー $D(\rho_t||\pi_\infty)$ の線形収束を示せ．（ρ_t は π_∞ に対して絶対連続であるとして良い．また，細かい正則条件は検証しないで良い）

問 10 (optional)　あるベクトル $a = (a_i)_i$ に対し，$\|a\|_p := (\sum_i |a_i|^p)^{1/p}$ $(1 \le p < \infty)$, $\|a\|_\infty := \max_i |a_i|$ とする．$(\sigma_i)_{i=1}^n$ を i.i.d. の Rademacher 確率変数とする; $P(\sigma_i = 1) = P(\sigma_i = -1) = 1/2$. また，$(x_i)_{i=1}^n$ は i.i.d. である確率分布から生成される確率変数であるとして，$\|x_i\|_\infty < 1$ (a.s.) を満たすものとする．ある関数クラス $\mathcal{F}$ の Rademacher 複雑度は

$$R_n(\mathcal{F}) = \mathrm{E}_{(x_i)_{i=1}^n,(\sigma_i)_{i=1}^n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i)\right].$$

で定義される．このとき，二層ニューラルネットワークの集合 $\mathcal{F}$ の Rademacher 複雑度を上から評価せよ:

$$\mathcal{F} = \left\{ \sum_{j=1}^M a_j \eta(w_j^\top x) \mid a_j \in \mathbb{R},\ w_j \in \mathbb{R}^d,\ \max_{1\le j\le M} \|w_j\|_p \le C_1,\ \|a\|_q \le C_2 \right\}.$$

ただし，$1 \le p, q \le \infty$ かつ $\eta(u) = \max\{u, 0\}$ とする．なお，1-Lipschitz 連続な関数 g に対して，$g \circ \mathcal{F} = \{g \circ f \mid f \in \mathcal{F}\}$ と書くとき，$R_n(g \circ \mathcal{F}) \le R_n(\mathcal{F})$ が満たされることは使っても良い．

参考文献

[1] D. Dowson and B. Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

[2] I. Karatzas and S. E. Shreve. *Brownian Motion and Stochastic Calculus, Second Edition.* Springer-Verlag New York, 1998.

[3] G. Obozinski and F. Bach. A unified perspective on convex structured sparsity: Hierarchical, symmetric, submodular norms and beyond. Preprint: hal-01412385, 2016.

[4] I. Olkin and F. Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263, 1982.

[5] C. Villani. *Topics in Optimal Transportation.* Graduate studies in mathematics. American Mathematical Society, 2003.