KDE 近似

师清

FDU

2023 年 4 月 7 日

Overview

- 问题定义
- paper: 《large-scale nonconvex optimization:randomization,gap estimation and numerical resolution》
  - Relaxation by randomization
  - Selection method
  - Frank-Wolfe Algorithm

问题定义:kde 近似

$$f(x) = \sum_{i=1}^{N_1} w_i^{'} exp\left[-\frac{(x-p_i)^2}{h}\right]$$

$$x_j \sim f(x) \rightarrow \{x_j\}_M$$

$$g(x) = \sum_{i=1}^{N_2} w_i exp\left[-\frac{(x-q_i)^2}{h}\right]$$

问题定义:kde 近似

$$
\begin{aligned}
&\|f(x)-g(x)\|_2^2 \\
&\approx \sum_{j=1}^{M}\left(f(x_j)-g(x_j)\right)^2 \\
&=\sum_{j=1}^{M}\left(t_j-g(x_j)\right)^2 \\
&=\sum_{j=1}^{j=M}\left\{\sum_{i=1}^{N} w_i exp\left[-\frac{(x_j-q_i)^2}{h}\right]-t_j\right\}^2 \\
&=\sum_{j=1}^{j=M}\left\{\sum_{i=1}^{N} w_i exp\left[-\frac{(y_j-x_i)^2}{h}\right]-t_j\right\}^2
\end{aligned}
$$

$$inf_{x\in\mathcal{X}} J(x) := f(G(x)) = f(\frac{1}{N}\sum_{i=1}^{N} g_i(x_i)) = \sum_{j=1}^{M} f_j(\frac{1}{N}\sum_{i=1}^{N} g_{ij}(x_i))$$

$$inf_{\mu\in P_\delta} \mathcal{J}(x) := f(\frac{1}{N}\sum_{i=1}^{N} E_{\mu_i}[g_i]) = \sum_{j=1}^{M} f_j(\frac{1}{N}\sum_{i=1}^{N} E_{\mu_i}[g_{ij}])$$

- Relaxation by randomization:repalce each optimization variable x_i by a probability measure μ_i, $g_i(x_i)$ are replaced by their integral with respect to μ_i, $E_{\mu_i}[g_i]$.
- Selection method: simulate N independent random variables $(X_1 ... X_N)$, with distributions $X_i \sim \mu_i$
- Frank-wolfe algorithm:solve randomized problem

Notations

$$S_{ij} := \{g_{ij}(x_i) | x_i \in \mathcal{X}\} \quad and \quad S_j := \frac{1}{N} S_{ij}$$

$$C_1 = \sum_{j=1}^{M} \left(L_j \max_{1 \le i \le N} \{d_{ij}\} \right) \quad and \quad C_1 = \frac{1}{N} \sum_{j=1}^{M} \left(\widetilde{L}_j \max_{1 \le i \le N} \{d_{ij}\} \right)$$

Assumption A. *For $i = 1, 2, \ldots, N$ and $j = 1, 2 \ldots, M$:*

1. *The range set S_{ij} in $\mathcal{E}_j$ has finite diameter $d_{ij} := d(S_{ij})$.*
2. *The function f_j is L_j-Lipschitz on* $\operatorname{conv}(S_j)$.
3. *The function f_j is continuously differentiable on a neighborhood of* $\operatorname{conv}(S_j)$, *and ∇f_j is $\tilde{L}_j$–Lipschitz on* $\operatorname{conv}(S_j)$, *in the sense of* (1.5).

Relaxation by randomization

LEMMA 2.5. *Let Assumption A hold true. Then* $-\infty < \mathcal{J}^* \leq J^*$.

Proof. By the definitions of $E_{\mu_i}[g_{ij}]$ and S_j, we have that $\frac{1}{N}\sum_{i=1}^{N} E_{\mu_i}[g_{ij}] \in \operatorname{conv}(S_j)$. Since f_j is Lipschitz-continuous over the bounded set $\operatorname{conv}(S_j)$, we deduce that $\mathcal{J}^* > -\infty$. Let $x \in \mathcal{X}$. Define $\mu = (\delta_{x_1}, \ldots, \delta_{x_N}) \in \mathcal{P}_\delta$. Then $\mathcal{J}(\mu) = J(x)$. As a consequence, inequality $\mathcal{J}^* \leq J^*$ follows. □

The *randomization gap* is then defined as

$$\text{randomization gap} = J^* - \mathcal{J}^* \geq 0.$$

Relaxation by randomization

Assumption B. *For all $j = 1, \ldots, M$, the function $f_j : \mathcal{E}_j \to \mathbb{R}$ is convex over* $\mathrm{conv}(S_j)$.

Let μ^1 and μ^2 lie in $\mathcal{P}_\delta$. Take $\omega \in [0, 1]$. Let $\mu = (\mu_1, \ldots, \mu_N)$ be defined, for any $i = 1, \ldots, N$, by $\mu_i = (1-\omega)\mu_i^1 + \omega\mu_i^2$. Here, the addition and the multiplication by a scalar are understood as usual in the set of signed measures. In the sequel, we simply denote $\mu = (1-\omega)\mu^1 + \omega\mu^2$. We have $\mu \in \mathcal{P}_\delta$; moreover, $E_{\mu_i}[g_i] = (1-\omega)E_{\mu_i^1}[g_i] + \omega E_{\mu_i^2}[g_i]$, for any $i = 1, \ldots, N$. Then, Assumption B implies that $\mathcal{J}(\mu) \leq (1-\omega)\mathcal{J}(\mu^1) + \omega\mathcal{J}(\mu^2)$. In words, the randomized problem (PR) is convex.

Relaxation by randomization

Proposition 2.6. *Let Assumption A hold true. Let $\mu \in \mathcal{P}_\delta$ and let $(X_i)_{i=1,\dots,N}$ denote N independent random variables such that $X_i \sim \mu_i$. Then,*

$$\mathbb{E}[J(X)] - \mathcal{J}(\mu) \leq \frac{1}{2N^2} \sum_{j=1}^{M} \left(\tilde{L}_j \sum_{i=1}^{N} \sigma^2_{\mu_i}[g_{ij}] \right) \leq \frac{C_1}{2N}, \tag{2.2}$$

where $X = (X_1, \dots, X_N)$. As a consequence, $J^* - \mathcal{J}^* \leq \frac{C_1}{2N}$.

Proof. Let us define $Y_j = \frac{1}{N}\left(\sum_{i=1}^{N} g_{ij}(X_i)\right)$, for $j = 1, \dots, M$. Let us set $Y = (Y_j)_{j=1,\dots,M}$. We have

$$\mathbb{E}\big[J(X)\big] = \mathbb{E}\big[f(Y)\big] \quad \text{and} \quad \mathcal{J}(\mu) = f\big(\mathbb{E}[Y]\big).$$

Since the variables X_i are independent, the random variables $g_{ij}(X_i)$ are also independent (for fixed j). It follows that

$$\mathbb{E}\Big[\|Y_j - \mathbb{E}[Y_j]\|^2_{\mathcal{E}_j}\Big] = \frac{1}{N^2} \sum_{i=1}^{N} \mathbb{E}\Big[\|g_{ij}(X_i) - \mathbb{E}[g_{ij}(X_i)]\|^2_{\mathcal{E}_j}\Big] = \frac{1}{N^2} \sum_{i=1}^{N} \sigma^2_{\mu_i}[g_{ij}].$$

By Assumption A, we have

$$f(Y) \leq f\big(\mathbb{E}[Y]\big) + \langle \nabla f(\mathbb{E}[Y]), Y - \mathbb{E}[Y] \rangle_{\mathcal{E}_j} + \frac{1}{2} \sum_{j=1}^{M} \Big(\tilde{L}_j \|Y_j - \mathbb{E}[Y_j]\|^2_{\mathcal{E}_j} \Big).$$

Taking the expectation of the above inequality and recalling the definition of C_1, we deduce (2.2). ∎

Relaxation by randomization

LEMMA 2.8. *Suppose that there exists a set* $\mathcal{X}$ *and a function* $g : \mathcal{X} \to \mathcal{E}$ *such that* $\mathcal{X}_i = \mathcal{X}$ *and* $g_i = g$, *for all* i. *Then,*

$$\mathcal{J}^* = \inf_{\nu \in \mathcal{P}_\delta(\mathcal{X})} f\big(E_\nu[g]\big). \tag{2.3}$$

Proof. Let $\nu \in \mathcal{P}_\delta(\mathcal{X})$. Take $\mu = (\nu, \dots, \nu) \in \mathcal{P}_\delta$. It follows that $f\left(E_\nu[g]\right) = \mathcal{J}(\mu)$. As a consequence, $\inf_{\nu \in \mathcal{P}_\delta(\mathcal{X})} f\left(E_\nu[g]\right) \leq \inf_{\mu \in \mathcal{P}_\delta} \mathcal{J}(\mu)$. On the other hand, let $\bar{\mu} = (\bar{\mu}_1, \dots, \bar{\mu}_N) \in \mathcal{P}_\delta$. Take $\bar{\nu} = \sum_{i=1}^N \bar{\mu}_i / N \in \mathcal{P}_\delta(\mathcal{X})$. Then, we deduce that $\mathcal{J}(\bar{\mu}) = f\left(E_{\bar{\nu}}[g]\right)$. The conclusion follows. $\square$

麦克迪尔米德不等式（McDiarmid's inequality）是机器学习中很重要的一个不等式，对于关于独立随机变量的函数的样本值与期望值的偏离，它给出了界。此不等式成立的条件是有界差性质（bounded difference property），即当我们只改变多元函数的一个变量时，函数值的差不能太大。对麦克迪尔米德不等式的证明用到了吾妻不等式。

定理 1.1（麦克迪尔米德不等式） 令 $S = (X_1, \cdots, X_n) \in \mathcal{X}^n$ *为一组独立随机变量. 假设存在常数* $c_1, \cdots, c_n > 0$ *和函数* $f : \mathcal{X}^n \to \mathbb{R}$ *，使得* $|f(x_1, \cdots, x_i, \cdots, x_n) - f(x_1, \cdots, x_i', \cdots, x_n)| \le c_i$ *，这对于所有* $i \in \{1, \cdots, n\}$ *和* $x_1, \cdots, x_n, x_i' \in \mathcal{X}^n$ *都成立. 那么，对于任意* $\epsilon > 0$ *，*

$$\mathbb{P}(f(S) - \mathbb{E}[f(S)] \ge \epsilon) \le \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

$$\mathbb{P}(f(S) - \mathbb{E}[f(S)] \le -\epsilon) \le \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

Selection method

Lemma 2.3. *Let Assumption A be satisfied. For all* $i \in \{1, \ldots, N\}$, *for all* $x_{-i} \in \mathcal{X}_{-i}$, x_i *and* x_i' *in* $\mathcal{X}_i$, *it holds:*

$$|J(x_i', x_{-i}) - J(x_i, x_{-i})| \leq \frac{C_0}{N}.$$

THEOREM 2.10. *Let Assumption A be satisfied. Let $\mu \in \mathcal{P}_\delta$ and let $X_1, \ldots, X_N$ be N independent random variables such that $X_i \sim \mu_i$. Let $X = (X_1, \ldots, X_N)$. Then, for all $\epsilon > 0$,*

$$\mathbb{P}\left[J(X) < \mathcal{J}(\mu) + \frac{C_1}{2N} + \epsilon\right] \geq 1 - \exp\left(-\frac{2N\epsilon^2}{C_0^2}\right). \tag{2.6}$$

Assume further that for all $i = 1, \ldots, N$, there exists a constant v_i such that

$$\sigma^2_{\mu_i}\left[J(\cdot, x_{-i})\right] \leq v_i^2, \tag{2.7}$$

for all $x_{-i} \in X_{-i}$. Then (2.6) can be strengthened as:

(2.8)

$$\mathbb{P}\left[J(X) < \mathcal{J}(\mu) + \sum_{j=1}^{M}\sum_{i=1}^{N} \frac{\tilde{L}_j}{2N^2}\sigma^2_{\mu_i}[g_{ij}] + \epsilon\right] \geq 1 - \exp\left(-\frac{N\epsilon^2}{2\left(\sum_{i=1}^{N} N v_i^2 + \frac{C_0\epsilon}{3}\right)}\right).$$

Proof. Combining Lemma 2.3 and McDiarmid's inequality [35], we obtain

$$\mathbb{P}\Big[J(X) < \mathbb{E}\big[J(X)\big] + \epsilon\Big] \geq 1 - \exp\left(-\frac{2N\epsilon^2}{C_0^2}\right).$$

Combining this estimate with the second inequality of Proposition 2.6, we obtain (2.6).

Estimate (2.8) is proved similarly, combining McDiarmid's inequality of "variance type" proved in Corollary A.2 and the first inequality of Proposition 2.6. □

Selection method

LEMMA 2.11. *Inequality* (2.7) *is satisfied with* $v_i^2 = \frac{1}{N^2}\left(\sum_{j=1}^M L_j^2\right)\sigma_{\mu_i}^2(g_i)$.

Proof. We first state a general following property: given a probability measure μ and two maps h_1 and h_2 suitably defined, we have the inequality

$$\sigma_\mu^2[h_1 \circ h_2] \le L^2 \sigma_\mu^2[h_2], \tag{2.9}$$

assuming that h_1 is L-Lipschitz continuous. Let us prove this property. For any x, we have

$$\begin{aligned}\left\|h_1 \circ h_2(x) - E_\mu[h_1 \circ h_2]\right\|^2 &= \left\|h_1 \circ h_2(x) - h_1(E_\mu[h_2])\right\|^2 \\ &\quad + 2\left\langle h_1 \circ h_2(x) - h_1(E_\mu[h_2]), h_1(E_\mu[h_2]) - E_\mu[h_1 \circ h_2]\right\rangle \\ &\quad + \left\|h_1(E_\mu[h_2]) - E_\mu[h_1 \circ h_2]\right\|^2.\end{aligned}$$

Taking the expectation, we obtain that

$$\sigma_\mu^2[h_1 \circ h_2] = E_\mu\left[\left\|h_1 \circ h_2 - h_1(E_\mu[h_2])\right\|^2\right] - \left\|h_1(E_\mu[h_2]) - E_\mu[h_1 \circ h_2]\right\|^2.$$

Since h_1 is L-Lipschitz continuous, we have $E_\mu\left[\|h_1 \circ h_2 - h_1(E_\mu[h_2])\|^2\right] \le L^2 \sigma_\mu[h_2]^2$. Inequality (2.9) follows immediately. Next, it is easy to verify that the function f is L-Lipschitz continuous, with $L = \left(\sum_{j=1}^M L_j^2\right)^{1/2}$. Using (2.9), we conclude that

$$\sigma_{\mu_i}^2[J(\cdot, x_{-i})] \le L^2 \sigma_{\mu_i}^2\left[\frac{1}{N} g_i(\cdot) + C\right] = \frac{L^2}{N^2}\sigma_{\mu_i}^2[g_i],$$

where $C = \frac{1}{N}\sum_{i' \neq i} g_{i'}(x_{i'})$ is regarded as a constant. The estimate follows. □

Frank-Wolfe algorithm

$$Define : \mathcal{A} := \{\nabla f(y) | y \in conv(G(\mathcal{X}))\}$$

Assumption C. *For all $i = 1, \ldots, N$, for all $\lambda \in \mathcal{A}$, the problem*

$$\inf_{x_i \in \mathcal{X}_i} \langle \lambda, g_i(x_i) \rangle \tag{3.1}$$

has at least a solution. For all $i = 1, \ldots, N$, we fix a map $\mathbb{S}_i : \mathcal{A} \mapsto \mathcal{X}_i$ such that for any $\lambda \in \mathcal{A}$, $\mathbb{S}_i(\lambda)$ is a solution to (3.1).

Lemma 3.2. *Let $\lambda \in \mathcal{A}$ and let $\bar{\mu} = (\bar{\mu}_1, \ldots, \bar{\mu}_N) \in \mathcal{P}_\delta$. Then, $\bar{\mu}$ is a solution to*

$$\inf_{\mu \in \mathcal{P}_\delta} \left\langle \lambda, \frac{1}{N} \sum_{i=1}^{N} E_{\mu_i}(g_i) \right\rangle. \tag{3.2}$$

if and only if for all $i = 1, \ldots, N$, $\bar{\mu}_i$ is supported in $\operatorname{argmin}_{x_i \in \mathcal{X}_i} \langle \lambda, g_i(x_i) \rangle$.

Frank-Wolfe algorithm

Algorithm 1 Frank-Wolfe Algorithm

Initialization: $\mu^0 \in \mathcal{P}_\delta$.
for $k = 0, 1, \ldots, K$ **do**
  Step 1: Resolution of the subproblems.
  Set $y^k = \frac{1}{N}\sum_{i=1}^{N} E_{\mu_i^k}[g_i]$ and set $\lambda^k = \nabla f(y^k)$.
  for $i = 1, \ldots, N$ **do**
    Compute $\bar{x}_i^k = \mathbb{S}_i(\lambda^k)$.
  end for
  Set $\bar{\mu}^k = (\delta_{\bar{x}_1^k}, \ldots, \delta_{\bar{x}_N^k})$.
  Step 2: Update.
  Set $\omega_k = 2/(k+2)$.
  Set $\mu^{k+1} = (1 - \omega_k)\mu^k + \omega_k \bar{\mu}^k$.
end for

Frank-Wolfe algorithm

$$(3.3)\quad \gamma_k = \mathcal{J}(\mu^k) - \mathcal{J}^*, \quad \beta_k = \langle \nabla f(y^k), y^k - \bar{y}^k \rangle, \quad \text{where:} \quad \bar{y}^k = \frac{1}{N}\sum_{i=1}^{N} g_i(\bar{x}_i^k).$$

LEMMA 3.3. *For all $k \in \mathbb{N}$, $\gamma_k \leq \beta_k$.*

Proof. Let $k \in \mathbb{N}$. Let $\mu \in \mathcal{P}_\delta$ and let $y = \frac{1}{N}\sum_{i=1}^{N} E_{\mu_i}[g_i]$. By Lemma 3.2, we have $\langle \nabla f(y^k), \bar{y}^k \rangle \leq \langle \nabla f(y^k), y \rangle$. Thus, using the convexity of f, we obtain

$$(3.4)\quad \beta_k = \langle \nabla f(y^k), y^k - \bar{y}^k \rangle \geq \langle \nabla f(y^k), y^k - y \rangle \geq f(y^k) - f(y) = \mathcal{J}(\mu^k) - \mathcal{J}(\mu).$$

Since μ is arbitrary, we deduce that $\beta_k \geq \mathcal{J}(\mu^k) - \mathcal{J}^* = \gamma_k$. $\square$

PROPOSITION 3.4. *Let Assumptions A, B, and C hold. Then, in Algorithm 1, for any $K \in \mathbb{N}^*$,*

$$\gamma_K \leq \frac{2C_1}{K}.$$

Proof. As we will see, the result is a consequence of Lemma A.3, with $C = \frac{C_1}{2}$ and $u_k = 0$. By Assumption A,

$$f(y^{k+1}) \leq f(y^k) + \langle \nabla f(y^k), y^{k+1} - y^k \rangle + \sum_{j=1}^{M} \frac{\tilde{L}_j}{2} \|y_j^{k+1} - y_j^k\|^2.$$

We have $y^{k+1} - y^k = \omega_k(\bar{y}^k - y^k)$. Therefore, by definition of β_k,

$$f(y^{k+1}) \leq f(y^k) - \omega_k \beta_k + \omega_k^2 \sum_{j=1}^{M} \frac{\tilde{L}_j}{2} \|\bar{y}_j^k - y_j^k\|^2. \tag{3.5}$$

By definition, $\|\bar{y}_j^k - y_j^k\|^2 = \frac{1}{N^2} \left\| \sum_{i=1}^{N} E_{\mu_i^k} \left[g_{ij}(\bar{x}_i^k) - g_{ij}(\cdot) \right] \right\|^2$, thus by Cauchy-Schwarz inequality,

$$\|\bar{y}_j^k - y_j^k\|^2 \leq \frac{1}{N} \sum_{i=1}^{N} \left\| E_{\mu_i^k} \left[g_{ij}(\bar{x}_i^k) - g_{ij}(\cdot) \right] \right\|^2 \leq \frac{1}{N} \sum_{i=1}^{N} d_{ij}^2.$$

Combining the above estimate with (3.5) and using the inequality $\gamma_k \leq \beta_k$ proved in Lemma 3.3, we obtain that $\gamma_{k+1} \leq (1 - \omega_k)\gamma_k + \frac{C_1}{2}\omega_k^2$. Thus Lemma A.3 applies, which concludes the proof. □

LEMMA A.3. *For all $k \in \mathbb{N}$, denote $\omega_k = \frac{2}{k+2}$. Let $(u_k)_{k\in\mathbb{N}}$ and $(\gamma_k)_{k\in\mathbb{N}}$ be two sequences of real numbers. Assume that there exists a positive number C such that*

$$\gamma_{k+1} \leq (1-\omega_k)\gamma_k + C\omega_k^2 + u_k, \tag{A.4}$$

for all $k \in \mathbb{N}$. Then, for all $K \in \mathbb{N}^*$,

$$\gamma_K \leq \frac{4C}{K} + \sum_{k=0}^{K-1} \frac{(k+1)(k+2)}{K(K+1)} u_k. \tag{A.5}$$

Proof. We proof this lemma by induction on K. We have $\omega_0 = 1$, thus taking $k = 0$ in (A.4), we obtain that $\gamma_1 \leq C + u_0$, which proves the claim for $K = 1$. Let us assume that the claim holds true for some $K \in \mathbb{N}^*$. We deduce from (A.4) that

$$\begin{aligned}\gamma_{K+1} &\leq \left(\frac{1}{K+2} + \frac{1}{(K+2)^2}\right)4C + \frac{K}{K+2}\left(\sum_{k=0}^{K-1} \frac{(k+1)(k+2)}{K(K+1)} u_k\right) + u_K \\ &\leq \frac{4C}{K+1} + \sum_{k=0}^{K} \frac{(k+1)(k+2)}{(K+1)(K+2)} u_k.\end{aligned}$$

Therefore the claim holds for $K+1$. This concludes the proof. ∎

Frank-Wolfe algorithm

Remark 3.5. For any $k \in \mathbb{N}$, denote $h_k(\omega) = -\omega\beta_k + \frac{C_k}{2}\omega^2$, where the constant C_k is defined by $C_k = \sum_{j=1}^{M} \tilde{L}_j \|\bar{y}_j^k - y_j^k\|^2$. In view of inequality (3.5), the result of Proposition 3.4 remains true if the sequence $(\omega_k)_{k\in\mathbb{N}}$ is chosen such that for any $k \in \mathbb{N}$, $h(\omega_k) \leq h(\bar{\omega}_k)$. The result remains in particular true for

$$\omega_k = \operatorname*{argmin}_{\omega\in[0,1]} h(\omega) = \min\left(\frac{\beta_k}{C_k}, 1\right). \tag{3.6}$$

Frank-Wolfe algorithm

LEMMA 3.6. *Let $(\mu_k)_{k\in\mathbb{N}}$ be the output of Algorithm 1. Let $k \le N$. Let $\zeta \in (0,1)$. Let $n \in \mathbb{N}^*$ and let $(X_i^j)_{i=1,\dots,N}^{j=1,\dots,n}$ be Nn independent random variables such that $X_i^j \sim \mu_i^k$. Let $X^j = (X_1^j, \dots, X_N^j)$. Then,*

$$\mathbb{P}\left[\min_{j=1,\dots,n} J(X^j) < \mathcal{J}^* + \frac{3C_1}{k}\right] \ge 1-\zeta, \quad \textit{if} \quad n \ge \frac{2C_0^2}{C_1^2}\frac{k^2}{N}\ln\left(\frac{1}{\zeta}\right). \tag{3.7}$$

Proof. Since $k \le N$, we have $\frac{C_1}{2N} \le \frac{C_1}{2k}$. Therefore, by Theorem 2.10,

$$\mathbb{P}\left[\min_{j=1,\dots,n} J(X^j) < \mathcal{J}^* + \frac{2C_1}{k} + \frac{C_1}{2k} + \epsilon\right] \ge 1 - \exp\left(-\frac{2N\epsilon^2 n}{C_0^2}\right),$$

for any $\epsilon > 0$. Take $\epsilon = \frac{C_1}{2k}$. If n satisfies (3.7), then $\exp\left(-\frac{2N\epsilon^2 n}{C_0^2}\right) \le \zeta$. □