Gradient Methods with Inexact Model in Clustering Problem

Olesya Kuznetsova
Optimization Class Project. MIPT

Introduction

In this project we consider gradient methods with inexact information of the objective given by inexact model of this objective. We analyze a gradient-type method for this type of problems and provide its convergence rate. To illustrate the applications, we consider optimization problems in a clustering model. Notably, our framework allows to solve non-convex problems which have a convex inexact model, which is illustrated in the section devoted to clustering model.

(δ, L)-model of a function

Definition 1. Let function $\psi_\delta(x, y)$ be convex in $x \in Q$ and satisfy $\psi_\delta(x, x) = 0$ for all $x \in Q$. We say that $\psi_\delta(x, y)$ is a (δ, L)-model of the function f in a given point y with respect to $V[y](x)$ iff for all $x \in Q$ the inequality

$$0 \leq f(x) - (f(y) + \psi_\delta(x, y)) \leq LV[y](x) + \delta$$

holds for some $L, \delta > 0$.

Such inexact model generalizes the concept of inexact oracle.

Definition 2. Consider a convex minimization problem

$$\phi(x) \to \min_{x \in Q \subseteq R^n}$$

If ϕ is smooth, we say that we solve it with $\tilde{\delta}$-precision ($\tilde{\delta} \geq 0$) if we find $\tilde{x}$ s.t. $\max_{x \in Q} \langle \nabla \phi(\tilde{x}), \tilde{x} - x \rangle = \tilde{\delta}$. If ϕ is general convex, we say that we solve this problem with $\tilde{\delta}$-precision if we find $\tilde{x}$ s.t. $\exists h \in \partial \phi(\widetilde{x}), \quad \langle h, x_* - \widetilde{x} \rangle \geq -\widetilde{\delta}$. In both cases we denote this $\tilde{x}$ as $\operatorname{argmin}^{\delta}_{x \in Q} \phi(x)$.

Algorithm

In this subsection we describe a gradient-type method for problems with (δ, L)-model of the objective.

1: **Input:** x_0 is the starting point, $L > 0$ and $\delta, \widetilde{\delta} > 0$.
2: **for** $k \geq 0$ **do**
3:

$$\phi_{k+1}(x) := \psi_\delta(x, x_k) + LV[x_k](x), \quad x_{k+1} := \arg\min^{\widetilde{\delta}}_{x \in Q} \phi_{k+1}(x). \quad (1)$$

4: **end for**
Output: $\bar{x}_N = \frac{1}{N} \sum_{k=0}^{N-1} x_{k+1}$

Convergence Rate

Theorem. Let $V[x_0](x_*) \leq R^2$, where x_0 is the starting point, and x_* is the nearest minimum point to the point x_0 in the sense of Bregman divergence $V[y](x)$. Then, for the sequence, generated by Algorithm the following inequality holds:

$$f(\bar{x}_N) - f(x_*) \leq \frac{LR^2}{N} + \widetilde{\delta} + \delta$$

Clustering Problem and Electoral Model

Consider so-called C-means soft clustering problem:

$$\min_{C \in R^{m \times K}} \left\{ \mathrm{Var}_\pi(C) = \sum_{i=1}^{N} \sum_{k=1}^{K} \pi_i^{(k)}(C) \, \|v_i - c_k\|_2^2 \right\}$$

, where C corresponds to the positions of the centers of the clusters and π are the probabilities of a cluster membership. Yu. Nesterov proposed the first soft-clustering model with the theoretically proved efficiency.

Let us describe an electoral model presented by Yu. Nesterov. In the model, we have N independent voters and K political parties and the main assumption is that the voting results are random. Voter i decides to vote for party k with probability p_i^k. We assume that an opinion of voter i can be described by a vector $v_i \in V$. At the same time, positions $x_k \in V$ are flexible. After each round of elections, these values can be adjusted for better representing the positions of the voters closely attached to this party.

Finally, let us fix some distance function $\rho(x, y)$ which is used for measuring the distance between the opinion of a voter and current position of a political party. In the electoral model, for probability distribution we apply the discrete choice probabilities of Logit model

$$p_i^{(k)}(X) = e^{-\rho(v_i, x_k)/\mu} \Big/ \left[\sum_{j=1}^{K} e^{-\rho(v_i, x_j)/\mu} \right], \quad k = 1, \ldots, K$$

where $\mu \geq 0$ is the flexibility parameter, which represents the volatility of opinions of voters.

In his paper, Yu. Nesterov shows that the process of clustering in the electoral model can be represented in a form of an optimization problem:

$$\min_{x_k \in V} \left\{ \hat{\psi}_k(P, x_k) = \frac{1}{N} \sum_{i=1}^{N} p_i^{(k)} \rho(v_i, x_k) + \frac{1}{\tilde{\tau}} d(c_k, x_k) \right\}$$

Solution of the Clustering Problem

So far, the problem can be rewritten as

$$f_{\mu_1, \mu_2}(x = (z, p)) = g(x) + \mu_1 \sum_{k=1}^{n} z_k \ln z_k + \frac{\mu_2}{2} \|p\|_2^2 \to \min_{z \in S_n(1), p \in {}^m_+}, \quad (2)$$

where ${}^m_+$ is a non-negative orthant and $S_n(1)$ is the standard n-dimensional simplex in ^{n}. In addition, we assume that $g(x)$ (generally, non-convex) is a function with L_g-Lipschitz continuous gradient:

$$\|\nabla g(x) - \nabla g(y)\|_* \leq L_g \|x - y\| \quad \forall x, y \in S_n(1) \times {}^m_+,$$

It can be shown that

$$\begin{aligned} \psi_\delta(x, y) = \langle \nabla g(y), x - y \rangle - L_g \cdot KL(z_x | z_y) - \frac{L_g}{2} \|p_x - p_y\|_2^2 \\ + \mu_1 (KL(z_x | \mathbf{1}) - KL(z_y | \mathbf{1})) + \frac{\mu_2}{2} \left(\|p_x\|_2^2 - \|p_y\|_2^2 \right) \end{aligned} \quad (3)$$

is a $(0, 2L_g)$-model of $f_{\mu_1, \mu_2}(x)$ in x with respect to the following Bregman divergence

$$V[y](x) = KL(z_x | z_y) + \frac{1}{2} \|p_x - p_y\|_2^2.$$

. Thus, the describe gradient method can be implemented to this problem.

Conclusion

We applied new method to, generally speaking, non-convex optimization problem which arises in clustering model. Derived convergence rate estimation is among the first theoretically proved estimations for this problem.