

Paper Presentation

Emil Joseph (ES15BTECH11009), Akhil Ahsref
(EE15BTECH11003)

IITH

March 4, 2019

PARALLEL STOCHASTIC SUCCESSIVE CONVEX APPROXIMATION METHOD FOR LARGE-SCALE DICTIONARY LEARNING

Alec Koppel, Aryan Mokhtari, and Alejandro Ribeiro

[Paper Link](#)

Abstract

Consider the problem of dictionary learning over training sets whose sample size and parameter dimension are large-scale, which is formulated as a non-convex stochastic program where the objective decomposes into a smooth non-convex part and a convex sparsity-promoting penalty.

This paper proposes a new method to find the optimum parameters from a non-convex objective function.

Setting up the problem

Consider a collection of signals $\mathbf{z}_n \in \mathbb{R}^p$. In dictionary learning, we have to find a corresponding $\alpha_n \in \mathbb{R}^k$ such that $\mathbf{z} = \alpha \mathbf{D}$ (in ideal case), where \mathbf{D} is the dictionary matrix $[d_1, \dots, d_k] \in \mathbb{R}^{p \times k}$. α is sparse and $k \geq p$.

The aim is to minimize the loss function,

$$(\mathbf{D}^*, \alpha^*) = \operatorname{argmin} \{E[\mathbf{D}\alpha - \mathbf{z}] + \lambda|\alpha|_1\} \quad (1)$$

Solution

To solve this problem, we iteratively find the best local optimum by converting the current set of data points by replacing the objective function with a convex surrogate function. So the objective function is re-written as,

$$V(x) := F(x) + \lambda|\alpha|_1 \quad (2)$$

where x is concatenation of \mathbf{D} and α

Assumptions on surrogate function $F(x)$

Assumption 1:

Consider x_i as the concatenation of all coordinates of x other than those of block i . The surrogate $f_i(x_i; x, z)$ associated with the i -th block of x , i.e., x_i , satisfies the following,

- 1) $f_i(x_i; x, z)$ is differentiable, convex w.r.t. x_i for all x, z .
- 2) $\nabla_{x_i} f_i(x_i; x, z) = \nabla_{x_i} f(x; z)$ for all x, z .
- 3) $\nabla_{x_i} f_i(x_i; x, z)$ is Lipschitz continuous on \mathcal{X} with constant Γ .

Assumption 2:

The sets χ_i are convex and compact.

Assumption 3:

Let F^t be the sigma-algebra generated by the collection of past realizations of x and z up to iteration t , i.e. $F^t \supset \{(x_u, z_u)\}_{u \leq t}$. The instantaneous gradients $\nabla_{x_i} f(x^t, z^t)$ induce stochastic errors whose conditional variance is finite:

$$\mathbf{E}[\|\nabla_{x_i} f(x^t, z^t) - \nabla_{x_i} F(x^t)\|^2 | F^t] \leq \sigma^2 \leq \infty \quad (3)$$

Assumption 4:

The statistical average objective $F(x)$ has L -Lipschitz continuous gradients, i.e., for all $x, y \in \mathcal{X}$

Doubly Stochastic Successive Convex approximation scheme (DSSC)

This is done on mini-batches which updates only a set of small number of parameters.

To do so, define the mini-batch sample surrogate function as,

$$f_i(x_i; x^t, \mathbf{z}_i^t) = \frac{1}{L} \sum_z f_i(x_i; x^t, z)$$

for a given a set of realizations \mathbf{z}_i^t , where L is the size of the mini-batch. Further define the mini-batch surrogate function gradient associated with the set

$$\mathbf{z}_i^t : \nabla f_i(x_i; x^t, \mathbf{z}_i^t) = \frac{1}{L} \sum_z \nabla f_i(x_i; x^t, z)$$

Cont.

With this, we can re-write the objective function as,

$$\begin{aligned}\hat{x}_i^{t+1} = \operatorname{argmin}_{x_i} \{ & \rho^t f_i(x_i; x^t, \mathbf{Z}_i^t) + (1 - \rho^t)(d_i^{t-1})^T (x_i - x_i^t) \\ & + \frac{\tau_i}{2} \|x_i - x_i^t\|^2 + g_i(x_i) \}\end{aligned}$$

where,

$$g_i(x_i) = \lambda \|\alpha_i\|_1 \quad (4)$$

Cont.

where,

$$d_i^t = (1 - \rho^t)(d_i^{t-1}) + \rho^t \nabla_{x_i} f_i(x_i; x^t, \mathbf{Z}_i^t) \quad (5)$$

x_i 's are updated as,

$$x_i^{t+1} = (1 - \gamma^{t+1})(x_i^t) + \gamma^{t+1} \hat{x}_i^{t+1} \quad (6)$$

where, γ is a known constant which is properly chosen.

Algorithm

Require: sequences γ^t and ρ^t .

for $t = 0, 1, 2, \dots$ **do**

Read the variable x_t .

Receive the randomly chosen block i

Choose training subset Z_i^t for block x_i

Compute surrogate function $f_i(x_i; x^t, \mathbf{Z}_i^t)$

Compute variable \hat{x}_i^{t+1} .

Compute surrogate gradient $\nabla f_i(x_i; x^t, \mathbf{Z}_i^t)$

Update average gradient d_i^t associated with block i

Compute the updated variable x_i^{t+1}

end

Parameter Selection

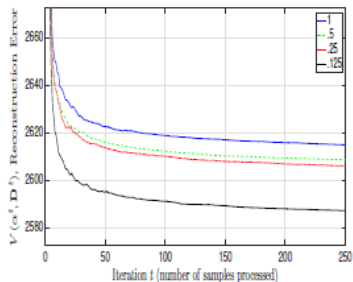
γ and ρ are chosen such that,

$$\lim_{t \rightarrow \infty} \gamma^t = 0, \sum_{t=0}^{\infty} \gamma^t = \infty, \sum_{t=0}^{\infty} (\gamma^t)^2 < \infty \quad (7)$$

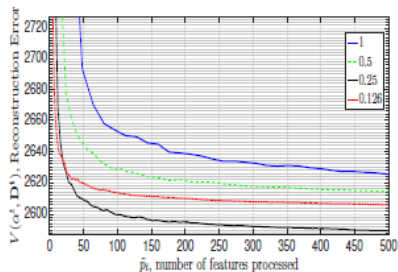
$$\lim_{t \rightarrow \infty} \rho^t = 0, \sum_{t=0}^{\infty} \rho^t = \infty, \sum_{t=0}^{\infty} (\rho^t)^2 < \infty \quad (8)$$

$$\sum_{t=0}^{\infty} \frac{(\gamma^t)^2}{\rho^t} < \infty \quad (9)$$

Their Results



(a) Objective $V(\mathbf{x}^t)$ vs. iteration t



(b) Objective $V(\mathbf{x}^t)$ vs. feature \tilde{p}_t

Figure 1: Results