Paper Presentation

Emil Joseph (ES15BTECH11009), Akhil Ahsref (EE15BTECH11003)

IITH

March 7, 2019

Topic

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 $z_n \epsilon R^p$. In dictionary learning, we have to find a corresponding $\alpha_n \epsilon R^k$ such that $\mathbf{z} = \alpha \mathbf{D}$ (in ideal case), where $\mathbf{D}$ is the dictionary matrix $[d_1, .., d_k] \epsilon R^{pxk}$. α is sparse and $k \geq p$.
The aim is to minimize the loss function,

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

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,

$$V(x) := F(x) + \lambda|\alpha|_1 \tag{2}$$

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 χ 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:

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

Assumption 4:
The statistical average objective F(x) has L-Lipschitz continuous gradients, i.e., for all $x, y \epsilon \chi$

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,

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

where,

$$g_i(x_i) = \lambda||\alpha_i||_1 \tag{4}$$

Cont.

where,

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

x_i's are updated as,

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

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

Proof of Convexity

$f_i(x_i; x^t, \mathbf{Z}_i^t)$ is the obtained convex surrogate function.

Proof that $(x_i - x_i^t)$ and $\lambda||\alpha_i||$ is convex (Note that x_i^t is a constant and $\alpha = \mathbf{D}^{-1}X$) :
Let $g(x) = (ax - c)$. where c is some constant vector.

$$\begin{aligned} g(\lambda x_1 + (1-\lambda)x_2) &= \lambda a x_1 + (1-\lambda)a x_2 - c \\ &= \lambda(a x_1 - c) + (1-\lambda)(a x_2 - c) \\ &= \lambda g(x_1) + (1-\lambda) g(x_2) \end{aligned}$$

Hence $(x_i - x_i^t)$ and and $\lambda||\alpha_i||$ is convex.

Proof of Convexity Cont.

Now to prove that $||x_i - x_i^t||^2$ is also convex.
Consider the function, $g(x) = ||x - c||^2$ where c is some constant vector.

$$\begin{aligned}
&\lambda g(x_1) + (1-\lambda)g(x_2) - g(\lambda x_1 + (1-\lambda)x_2) \\
&= \lambda||x_1 - c||^2 + (1-\lambda)||x_2 - c||^2 - ||\lambda x_1 + (1-\lambda)x_2 - c||^2 \\
&= \lambda(1-\lambda)\{||x_1 - c||^2 + ||x_2 - c||^2 - 2||x_1 - c||||x_2 - c||\} \\
&= \lambda(1-\lambda)\{||x_1 - c|| - ||x_1 - c||\}^2 \\
&\geq 0
\end{aligned}$$

Therefore,

$$\lambda g(x_1) + (1-\lambda)g(x_2) \geq g(\lambda x_1 + (1-\lambda)x_2)$$

Hence $||x_i - x_i^t||^2$ is convex.

Proof of Convexity Cont.

Our objective function is a conical (linear with coefficients ≥ 0) combination of the above convex functions. Hence the whole objective function is convex.

Algorithm

Require: sequences γ^t and ρ^t.
for t = 0,1,2... **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

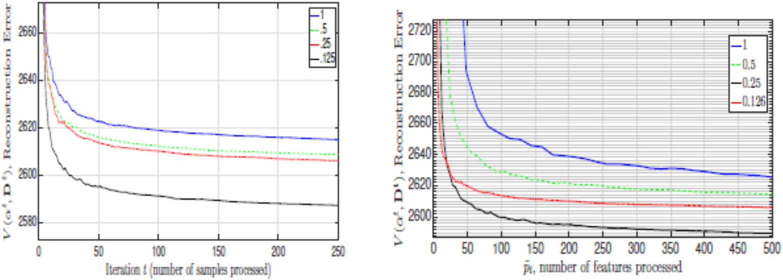
Parameter Selection

γ and ρ are chosen such that,

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

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

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



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

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

Figure 1: Results