Presentation in *Convex Optimization*

Mingrui Zhang, Xialiang Dou, Dongming Huang

Dec 22, 2014

Sample size selection in optimization methods for machine learning

Sample size selection in optimization methods for machine learning

Main results: presents a methodology for using varying sample sizes in batch-type optimization methods for large-scale machine learning problems.

Sample size selection in optimization methods for machine learning

Main results: presents a methodology for using varying sample sizes in batch-type optimization methods for large-scale machine learning problems.

- Dynamic sample selection in the evaluation of the function and gradient.
- A practical Newton method that uses smaller sample to compute Hessian vector-products.

The dynamic sample size gradient method

Problem: To determine the values of the parameters $\omega \in \mathbb{R}^m$ of a prediction function $f(\omega; x)$, where we assume:

$$f(\omega, x) = \omega^{\mathrm{T}} x \tag{1}$$

The dynamic sample size gradient method

Problem: To determine the values of the parameters $\omega \in \mathbb{R}^m$ of a prediction function $f(\omega; x)$, where we assume:

$$f(\omega, x) = \omega^{\mathrm{T}} x \tag{1}$$

Common Approach:To minimize the empirical loss function:

$$J(\omega) = \frac{1}{N} \sum_{i=1}^{N} l(f(\omega; x_i), y_i) \tag{2}$$

where $l(\hat{y}, y)$ is a convex loss function.

The dynamic sample size gradient method

Assumption: The size of the data set N is extremely large, numbered somewhere in the millions or billions, so that the evaluation of $J(\omega)$ is very expensive.

The dynamic sample size gradient method

Assumption: The size of the data set N is extremely large, numbered somewhere in the millions or billions, so that the evaluation of $J(\omega)$ is very expensive.

A gradient-based mini-batch optimization algorithm: At every iteration, chooses a subset $\mathcal{S} \subset \{1, 2, \cdots, N\}$ of the training set, and applies one step of an optimization algorithm to the objective function:

$$J_{\mathcal{S}}(\omega) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} l(f(\omega; x_i), y_i) \tag{3}$$

The dynamic sample size gradient method

Measure of the quality of the sample $\mathcal{S}$: the variance in the gradient $\nabla J_{\mathcal{S}}$

The dynamic sample size gradient method

Measure of the quality of the sample $\mathcal{S}$: the variance in the gradient $\nabla J_{\mathcal{S}}$

It's easy to verify that the vector $d = -\nabla J_{\mathcal{S}}(\omega)$ is a descent direction for J at ω if:

$$\delta_{\mathcal{S}}(\omega) \equiv \|\nabla J_{\mathcal{S}}(\omega) - \nabla J(\omega)\|_2 \leq \theta \|\nabla J_{\mathcal{S}}(\omega)\|_2 \tag{4}$$

where $\theta \in [0, 1)$.

Note that

$$\mathbb{E}[\delta_{\mathcal{S}}(\omega)^2] = \mathbb{E}[\|\nabla J_{\mathcal{S}}(\omega) - J(\omega)\|_2^2] = \|\mathrm{Var}(\nabla J_{\mathcal{S}})\|_1 \tag{5}$$

By simple calculations, we have:

$$\mathrm{Var}(\nabla J_{\mathcal{S}}(\omega)) = \frac{\mathrm{Var}(\nabla l(\omega; i))}{|\mathcal{S}|} \frac{N - |\mathcal{S}|}{N - 1} \tag{6}$$

where $\frac{N - |\mathcal{S}|}{N - 1} \approx 1$.

The dynamic sample size gradient method

By simple calculations, we have:

$$\mathrm{Var}(\nabla J_{\mathcal{S}}(\omega)) = \frac{\mathrm{Var}(\nabla l(\omega; i))}{|\mathcal{S}|} \frac{N - |\mathcal{S}|}{N - 1} \tag{6}$$

where $\dfrac{N - |\mathcal{S}|}{N - 1} \approx 1$.

Then we can rewrite the condition in the following format:

$$|\mathcal{S}| = \frac{\|\mathrm{Var}_{i \in \mathcal{S}}(\nabla l(\omega; i))\|_1}{\theta^2 \|\nabla J_{\mathcal{S}}(\omega)\|_2^2} \tag{7}$$

This is also the criterion that we use to determine the dynamic sample size.

A Newton-CG method with dynamic sampling

At each iteration, the subsampled Newton-CG method chooses samples $\mathcal{S}_k$ and $\mathcal{H}_k$ such that $|\mathcal{H}_k| \ll |\mathcal{S}_k|$, and defines the search direction d_k as an approximate solution of the linear system

$$\nabla^2 J_{\mathcal{H}_k}(\omega_k)d = -\nabla J_{\mathcal{S}_k}(\omega_k) \tag{8}$$

A Newton-CG method with dynamic sampling

At each iteration, the subsampled Newton-CG method chooses samples $\mathcal{S}_k$ and $\mathcal{H}_k$ such that $|\mathcal{H}_k| \ll |\mathcal{S}_k|$, and defines the search direction d_k as an approximate solution of the linear system

$$\nabla^2 J_{\mathcal{H}_k}(\omega_k)d = -\nabla J_{\mathcal{S}_k}(\omega_k) \tag{8}$$

Now we turn to create automatic criterion for deciding the accuracy in the solution of (8)

$$r_k \equiv \nabla^2 J_{\mathcal{H}_k}(\omega_k)d + \nabla J_{\mathcal{S}_k}(\omega_k) \tag{9}$$

$$r_k \equiv \nabla^2 J_{\mathcal{H}_k}(\omega_k)d + \nabla J_{\mathcal{S}_k}(\omega_k) \tag{9}$$

Then we write the residual of the standard Newton iteration as:

$$\nabla^2 J_{\mathcal{S}_k}(\omega_k)d + \nabla J_{\mathcal{S}_k}(\omega_k) = r_k + [\nabla^2 J_{\mathcal{S}_k}(\omega_k) - \nabla^2 J_{\mathcal{H}_k}(\omega_k)]d \tag{10}$$

$$r_k \equiv \nabla^2 J_{\mathcal{H}_k}(\omega_k)d + \nabla J_{\mathcal{S}_k}(\omega_k) \tag{9}$$

Then we write the residual of the standard Newton iteration as:

$$\nabla^2 J_{\mathcal{S}_k}(\omega_k)d + \nabla J_{\mathcal{S}_k}(\omega_k) = r_k + [\nabla^2 J_{\mathcal{S}_k}(\omega_k) - \nabla^2 J_{\mathcal{H}_k}(\omega_k)]d \tag{10}$$

If we define:

$$\mathbb{E}[\Delta_{\mathcal{H}_k}(\omega_k; d)^2] \equiv [\nabla^2 J_{\mathcal{S}_k}(\omega_k) - \nabla^2 J_{\mathcal{H}_k}(\omega_k)]d \tag{11}$$

Then we can make the approximation:

$$\mathbb{E}[\Delta_{\mathcal{H}_k}(\omega_k; d)^2] \approx \frac{\|\mathrm{Var}_{i\in\mathcal{H}_k}(\nabla^2 l(\omega_k; i)d)\|_1}{|\mathcal{H}_k|} \tag{12}$$

A Newton-CG method with dynamic sampling

In order to avoid to recompute the variance at every CG iteration, we initialize the CG iteration at the zero vector. From (9), we have that the initial CG search direction is given by
$p_0 = -r_0 = -\nabla J_{\mathcal{S}_k}(\omega_k)$.

A Newton-CG method with dynamic sampling

In order to avoid to recompute the variance at every CG iteration, we initialize the CG iteration at the zero vector. From (9), we have that the initial CG search direction is given by
$p_0 = -r_0 = -\nabla J_{\mathcal{S}_k}(\omega_k)$.

We compute (12) at the beginning of the CG iteration, for $d = p_0$.

A Newton-CG method with dynamic sampling

In order to avoid to recompute the variance at every CG iteration, we initialize the CG iteration at the zero vector. From (9), we have that the initial CG search direction is given by
$p_0 = -r_0 = -\nabla J_{\mathcal{S}_k}(\omega_k)$.

We compute (12) at the beginning of the CG iteration, for $d = p_0$.

The stop test for the $j+1$ CG iteration is then set as:

$$\|r_{j+1}\|_2^2 \leq \Psi \equiv \left(\frac{\|\mathrm{Var}_{i\in\mathcal{H}_k}(\nabla^2 l(\omega_k; i)p_0)\|_1}{|\mathcal{H}_k|}\right)\frac{\|d_j\|_2^2}{\|p_0\|_2^2} \tag{13}$$

where d_j is the jth trial candidate for the solution of (8) generated by the CG process and the last ratio accounts for the length of the CG solution.