Variance reduction in ellipsoid method

Andrey Filatov
Optimization Class Project. MIPT

Introduction

In the tasks of minimization of convex, but not necessarily smooth and strongly convex, the functions of a large number of summands cannot be obtained linear convergence using the gradient descent and gradient reduction methods. If the space dimension is small, the ellipsoid method can be used to obtain linear convergence. The calculation of the gradient of the total amount can be replaced by batching. The two approaches above lead to a method of variance reduction for the ellipsoidal method.

Notation

- X — given convex and compact set
- n — space dimension
- N — number of iterations
- c — center of ellipsoid
- r — radius of ball incircled of X
- R — radius of ball excircled of X
- ∇f — true gradient
- $\nabla^k f$ — batch gradient
- $\partial^k f$ — batch subgradient

Ellipsoid method

Let $f : X \to [-B, B]$ — continious convex function. X — convex compact, which is contained in the eucledian ball radius of R and is contained eucledian ball radius of r.

Algorithm 1: Ellipsoid method with batching

Input: Number of iteration N, ball $B_R \supseteq X$, c — center and R – radius
Output: $\tilde{x} \in \mathbb{R}^n$
$\mathcal{E}_0 \leftarrow \mathcal{B}_R, H_0 \leftarrow R^2 I_n, c_0 \leftarrow c$;
for $t = 0, \ldots, N-1$ **do**
  if $c_t \in X$ **then**
    $w_t \leftarrow w \in \partial^k f(c_t)$;
    if $w_t = 0$ **then**
      $\tilde{x} \leftarrow c_t$;
      return $\tilde{x}$
    end
  else
    $w_t \leftarrow w$, where $w \neq 0$, and $X \subset \left\{x \in \mathcal{E}_l : w^T (x - c_l) \leq 0\right\}$
  end
  $c_{t+1} \leftarrow c_t - \frac{1}{n+1} \frac{H_t w_t}{\sqrt{w_t^T H_t w_t}}$;
  $H_{t+1} \leftarrow \frac{n^2}{n^2-1} \left(H_t - \frac{2}{n+1} \frac{H_t w_t w_t^T H_t}{w_t^T H_t w_t}\right)$;
  $\mathcal{E}_{t+1} \leftarrow \left\{x : (x - c_{t+1})^T H_{t+1}^{-1} (x - c_{t+1}) \leq 1\right\}$
end
return $\tilde{x} = \underset{x \in \{c_0 \ldots c_N\} \cap X}{\operatorname{argmin}} f(x)$

Theorem

For $N \geq 2n^2 \ln \frac{R}{r}$ ellipsoid method with batching return $\tilde{x} \in X$, s.t.

$$f(\bar{x}) - f(x_*) \leq \frac{2BR}{r} \exp\left(-\frac{N}{2n^2}\right) + \delta$$

with the probability equals to $\left[F\left(\frac{\delta}{\sigma R}\right)\right]^N = \left[\int_0^{\frac{\delta}{\sigma R}} \frac{x^{n-1} e^{-x^2/2}}{2^{n/2-1}\Gamma(n/2)} dx\right]^N$

Theorem proof

Suppose the difference between the true gradient and the batching gradient $\nabla f - \nabla^k f$ is distributed as $\mathcal{N}(0, \sigma^2)$. Then euclidean norm of difference $\|\nabla f - \nabla^k f\|_2$ will have chi distribution multiplied by σ. To estimate δ for δ-subgradient we can use Cauchy-Schwarz inequality.

$$f(y) \geqslant f(x) + \langle \nabla f, y - x \rangle \wedge f(y) \geqslant f(x) + \langle \nabla f, y - x \rangle - \delta$$

$$\langle \nabla f - \nabla^k f, y - x \rangle \leqslant \|\nabla f - \nabla^k f\| \|R\| = \delta$$

As $\|\nabla f - \nabla^k f\|$ is a random variable to gain convergence rate we can use the proved theorem for constant δ-subgradient if we take $\max\limits_{1 \ldots N} \delta_k$.

Using our assumption we can estimate the probability of max on N iterations will be less than δ. Distribution of maximum is:

$$\left[F\left(\frac{\delta}{\sigma R}\right)\right]^N = \left[\int_0^{\frac{\delta}{\sigma R}} \frac{x^{n-1} e^{-x^2/2}}{2^{n/2-1}\Gamma(n/2)} dx\right]^N$$

In an equal way, we can estimate the expected value:

$$\mathbb{E}\delta = \sigma R N \int_0^{\infty} y \left[F(y)\right]^{N-1} \frac{y^{n-1} e^{-y^2/2}}{2^{n/2-1}\Gamma(n/2)} dy$$

Estimation of σ can be done through an assumption that the expected value of the square norm of difference between true gradient and one element gradient less than D:

$$\mathbb{E}\|\nabla f - \nabla^1 f\|_2^2 \leqslant D$$

Then according to batching, we have such estimate for the batch gradient

$$\mathbb{E}\|\nabla f - \nabla^k f\|_2^2 \leqslant \frac{D}{k}$$

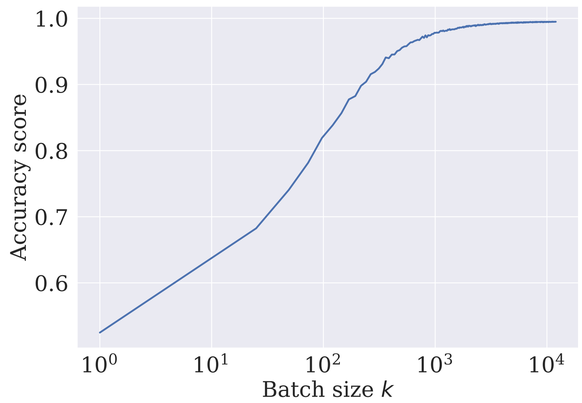
Distribution of squared gradient difference in our assumption will have chi-squared distribution multiplied by σ^2. Then the expected value of this distribution is $\sigma^2 n \leqslant \frac{D}{k}$ and we gain estimate for σ what completes the proof.

Numerical experiment

The proposed method was investigated on MNIST dataset. Two classes: "3" and "6" were chosen for a binary classification task. As loss function, we have chosen SVM loss + l2 regularization, because this task meets our conditions of convexity and smoothness. Data dimension was reducted from 784 to 100 through PCA[5]. After this, data was normalized. The initial ellipsoid was a ball with the centre in zero and the radius is equal to 5.0. The stopping criterium was H-norm of the gradient is less than $(0.1)^{-100}$. The code is available on github/anvilarth

Results

Results on every batch size were averaged over 30 iterations.



Conclusion

Proposed theoretical explanation of batching ellipsoid method. Gained estimate of convergence rate. We evaluate the batching method of SVM task and show practical applicability for batched ellipsoid method. In future works, we will try to improve this estimate or prove its imperfection.

References

[1] S. Bubeck. Convex optimization: Algorithms and complexity, 2014.

[2] A. Gasnikov. Universal gradient descent, 2017.

[3] Gladin. Ellipsoid method with an inexact subgradient. 2020.

[4] L. G. Khachiyan. A polynomial algorithm in linear programming. In *Doklady Akademii Nauk*, volume 244, pages 1093–1096. Russian Academy of Sciences, 1979.

[5] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.