Adaptive selection of the learning rate in Stochastic Gradient Methods

Giorgia Franchini, Valeria Ruggiero and Luca Zanni

Department of Physics, Informatics and Mathematics, University of Modena and Reggio Emilia, Modena (Italy)

UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

HiPeRT Lab High-Performance Real-Time Lab

CLASS www.class-project.eu

Stochastic Gradient Methods

- ▷ The problem we consider is the unconstrained minimization of the form

$$\min_x F(x) = \mathbb{E}[f(x,\xi)]$$

where ξ is a multi-value random variable and f represents the cost function. For example: minimize the sum of cost functions depending on a finite training set, composed by sample data ξ_i, $i \in \{1 \dots n\}$:

$$\min_x F_n(x) = \frac{1}{n}\sum_{i=1}^{n} f(x,\xi_i) = \frac{1}{n}\sum_{i=1}^{n} f_i(x), \qquad (1)$$

where n is the number of samples and each $f_i(x) \equiv f(x,\xi_i)$ denotes the cost function related to the instance ξ_i of the training set elements.

- ▷ When n is large, computing $F(x)$ and $\nabla F(x)$ is prohibited;
- ▷ Stochastic Gradient (SG) method and its variants have been the main approaches for solving (1);
- ▷ in the $t-$th iteration of SG, a random index of a training sample i_t is chosen from $\{1,2,\dots,n\}$ and the iterate x_t is updated by

$$x_{t+1} = x_t - \eta_t \nabla f_{i_t}(x_t)$$

where $\nabla f_{i_t}(x_t)$ denotes the gradient of the i_t-th component function at x_t, and $\eta_t > 0$ is the steplength or learning rate, [1].

Adaptive steplength selection in the stochastic framework

- **The deterministic framework: selections based on the Ritz-like values [2]**

Choose the steplengths for m_R next iterations as

$$\eta^R_{t-1+i} = \frac{1}{\theta_i}, \qquad i = 1,\dots,m_R \qquad (m_R = 3,4,5)$$

where θ_i are the eigenvalues of an $m_R \times m_R$ symmetric tridiagonal matrix T derived from the last m_R gradients

$$[\nabla F(x_{t-m_R}),\dots,\nabla F(x_{t-1})]$$

by generalizing the Lanczos process for approximating the eigenvalues of a symmetric matrix.

In case of quadratic objective function ($F(x) = \frac{1}{2}x^T Ax - b^t x$), the values θ_i (called *Ritz* values) are approximations of m_R eigenvalues of the symmetric positive definite matrix A.

In the general non-quadratic case, the values θ_i tend to approximate m_R eigenvalues of the Hessian matrix at the solution [3].

Compute the symmetric tridiagonal matrix T

- ▷ Let $\quad G = [\nabla F(x_{t-m_R}),\dots,\nabla F(x_{t-1})] \qquad (m_R = 3,4,5)$
- ▷ Compute the Cholesky decomposition $\quad G^T G = R^T R$ where $R_{m_R \times m_R}$ is upper triangular
- ▷ Compute

$$J = \begin{pmatrix} \eta^{-1}_{t-m_R} & & & \\ -\eta^{-1}_{t-m_R} & \ddots & & \\ & \ddots & & \eta^{-1}_{t-1} \\ & & & -\eta^{-1}_{t-1} \end{pmatrix}$$

- ▷ Compute $\tilde{T}$

$$\tilde{T} = [R \;\; \boldsymbol{v}]\, J\, R^{-1} \qquad \text{where} \qquad R^T \boldsymbol{v} = G^T \nabla F(x_t)$$

- ▷ T = *tril*(T) + *tril*(T,-1)'

- **The stochastic framework: Selection based on Ritz-like values in SG**

Exploit

$$\tilde{G} = [\nabla f_{t-m_R}(x_{t-m_R}),\dots,\nabla f_{t-1}(x_{t-1})]$$

in computing the *Ritz-like* values θ_i for the next m_R iterations and set in SGD

$$\eta_t = \max\left\{\min\left\{\eta_{max}, \frac{1}{\theta_i}\right\}, \eta_{min}\right\}$$

The test problem

- We built linear classifiers corresponding to three different loss functions (logistic regression, square loss, smooth hinge loss); in all cases, a regularization term was added to avoid overfitting. Thus the minimization problem has the form

$$\min_x F_n(x) + \frac{\lambda}{2}\|x\|_2^2,$$

where $\lambda > 0$ is a regularization parameter, $a_i \in \mathbb{R}^d$ and $b_i \in \{1,-1\}$ are the feature vector and the class label of the $i-th$ sample, respectively;

- The loss function $F_n(x)$ assumes one of the following form:
  - logistic regression:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \log\left[1 + e^{(-b_i a_i^T x)}\right];$$

  - square loss:

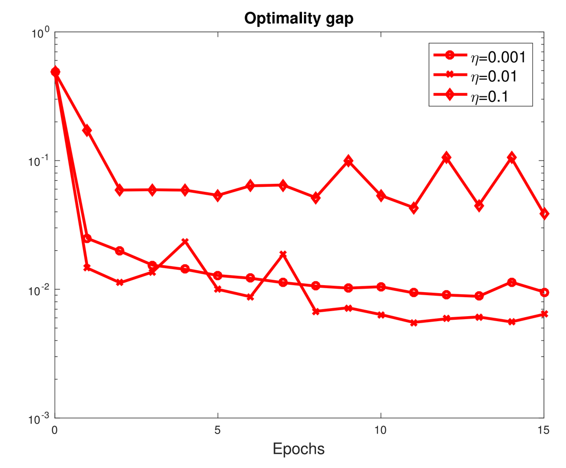
$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n}(1 - b_i a_i^T x)^2;$$

  - smooth hinge loss:

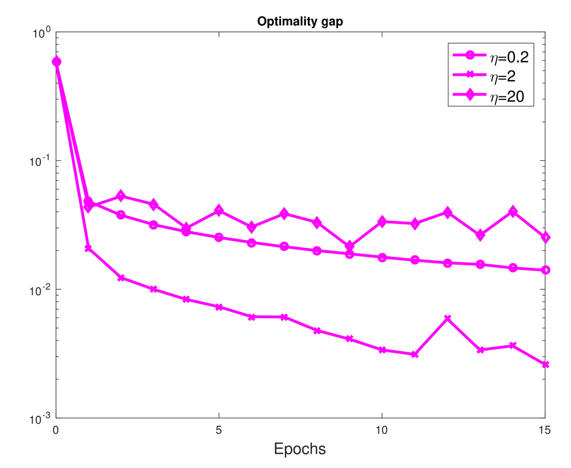
$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases} \frac{1}{2} - b_i a_i^T x, & \text{if } b_i a_i^T x \le 0 \\ \frac{1}{2}(1 - b_i a_i^T x)^2, & \text{if } 0 < b_i a_i^T x < 1 \\ 0, & \text{if } b_i a_i^T x \ge 1 \end{cases}$$

- We consider the two well known data-sets:
  - the *MNIST* data-set of handwritten digits, the images are in gray-scale $(0,255)$, in our case normalized $(0,1)$, centered in a box of 28×28 pixels; from the whole data-set of $60,000$ images, $11,800$ images were extracted exclusively relating to digits 8 and 9;
  - the web data-set *w8a* containing 49,749 examples; each example is described by 300 binary features.

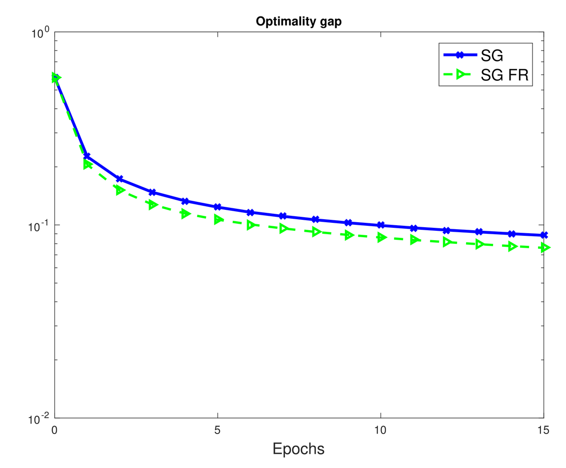
Experimental results



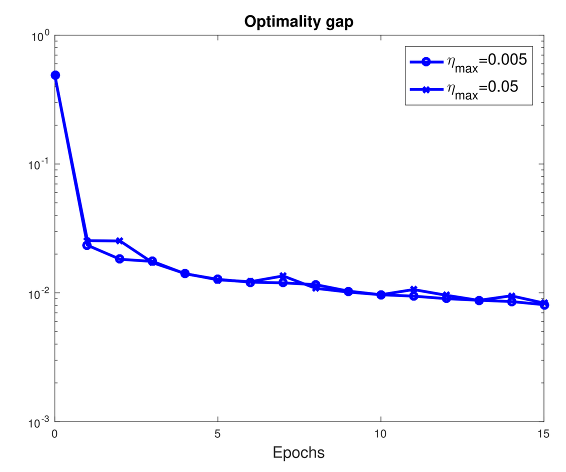
Behaviour of SG with different steplenghts over 15 epochs on the MNIST data set; test problem with smooth hinge loss function.



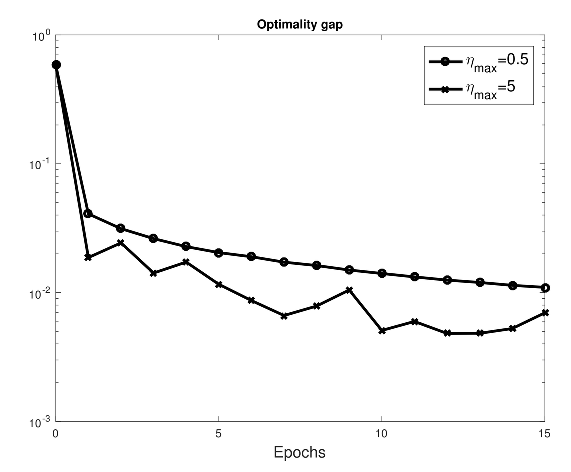
Behaviour of SG mini-batch with different steplenghts over 15 epochs on the w8a data set; test problem with logistic regression loss function.



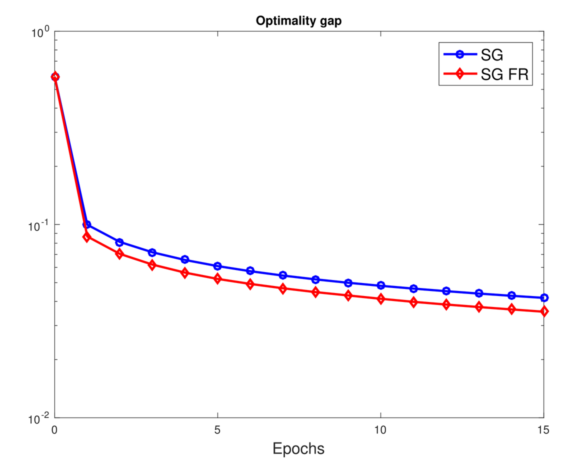
Behaviour of SG with $\eta = 0.0001$ and SG FR with $\eta_{max} = 0.0005$ over 15 epochs on the w8a data set; test problem with smooth hinge loss function.



Behaviour of SG FR with different η_{max} over 15 epochs on the MNIST data set; test problem with smooth hinge loss function.



Behaviour of SG FR mini-batch with different η_max over 15 epochs on the w8a data set; test problem with logistic regression loss function.



Behaviour of SG and SG FR with $|S| = 20; \eta = 0.02\eta_{max} = 0.05$ over 15 epochs on the w8a data set; test problem with smooth hinge loss function.

Conclusion and perspective work

- The proposed steplength approach depends on the chosen interval $[\eta_{min},\eta_{max}]$ and on η_{ini}, the effectiveness of the corresponding SG methods is slightly affected by variations of these parameters;
- This behaviour introduces greater flexibility with respect to the choice of a fixed small scalar, that must be carefully tuned;
- Future works will involve variance reduction methods and its validation on other loss functions;
- Following the suggestions of [Bollapragada et al. 2017], a very interesting analysis will concern the possibility of combining the proposed steplength selection rule with inexact line search techniques used in SG methods.

References

[1] L. Bottou, F.E. Curtis, J. Nocedal, *Optimization methods for large-scale machine learning*, SIAM Review, 2018 - SIAM
[2] R. Fletcher, *A limited memory steepest descent method*, Mathematical Programming, Volume 135, Springer (2012)
[3] D.Serafino, V.Ruggiero, G.Toraldo, L.Zanni, *On the steplength selection in gradient methods for unconstrained optimization*, Appl. Math. Comput. 318(2018) 176–195.
[4] Sopyla, Drozda, *SGD with BB update step for SVM*, Inf. Sci., 2015
[5] Tan, Ma, Dai, Qian, *BB Step Size for SGD*, Adv NIPS 2016