On the steplength selection in Stochastic Gradient Methods

Giorgia Franchini, and Luca Zanni

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

Stochastic Gradient Methods

- The following optimization problem, which minimizes the sum of cost functions over samples from a finite training set, appears frequently in machine learning

$$\min F(x) \equiv \frac{1}{n}\sum_{i=1}^{n} f_i(x), \qquad (1)$$

where n is the number of samples, and each $f_i : \mathbb{R}^d \to \mathbb{R}$ is the cost function corresponding to a training set element.

- 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 [4].

Compute the symmetric tridiagonal matrix T

- Let $G = [\nabla F(x_{t-m_R}), \dots, \nabla F(x_{t-1})] \qquad (m_R = 3, 4, 5)$
- Compute the Cholesky decomposition $G^T G = R^T R$ 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

$$\eta_t = \max\left\{\min\left\{10^2\eta_0, \frac{1}{\theta_i}\right\}, 10^{-1}\eta_0\right\}$$

The test problem

- Logistic regression with $l_2 - norm$ regularization:

$$\min_x F(x) = \frac{1}{n}\sum_{i=1}^{n} \log\left[1 + exp(-b_i a_i^T x)\right] + \frac{\lambda}{2}\|x\|_2^2$$

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

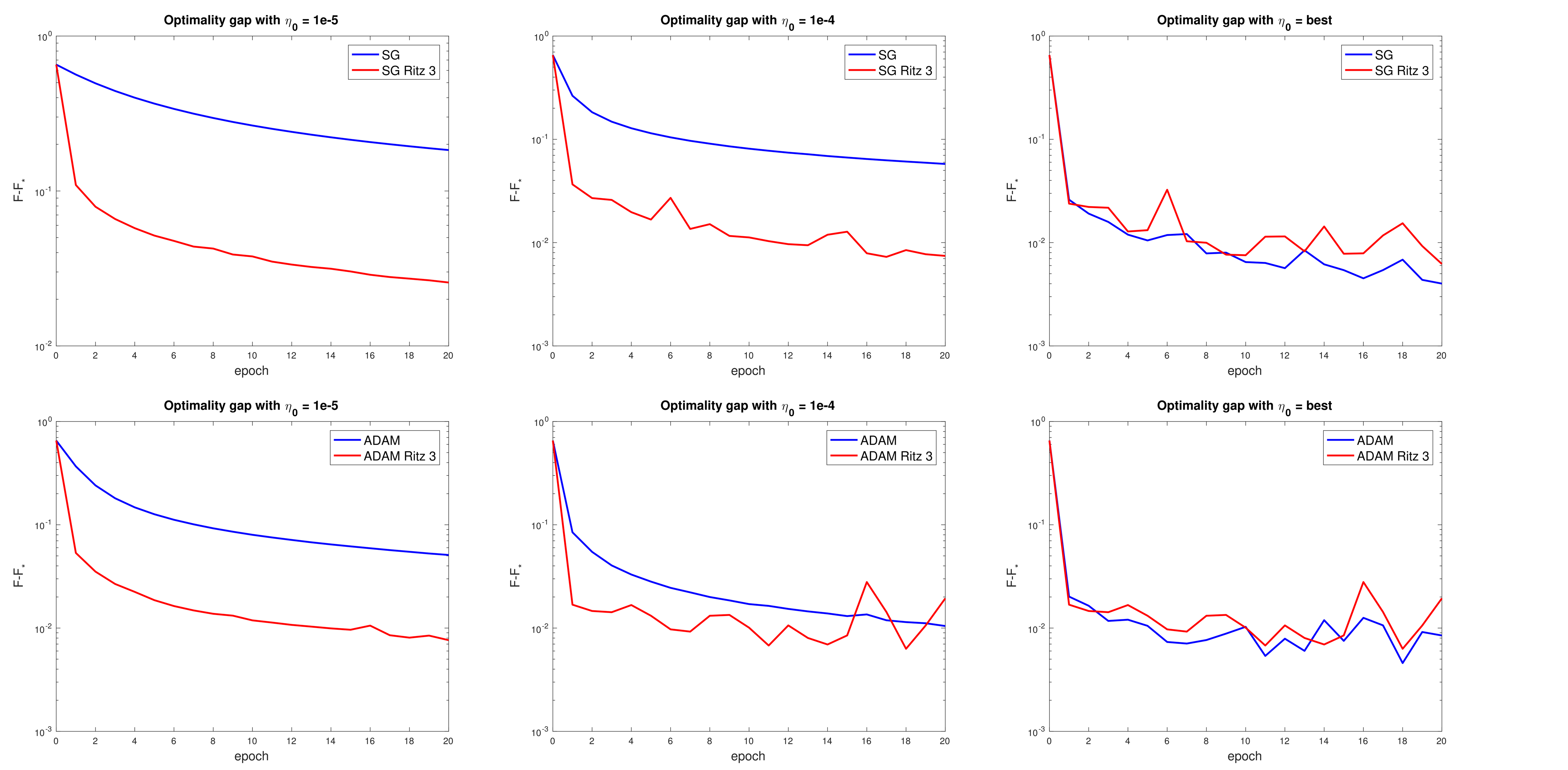
- database: MNIST 8 and 9 digits (binary classification), dimension: 11800×784.

ADAM algorithm [3]

Algorithm 1 Adam

```
Choose maxit, η, ε, β1 and β2 ∈ [0,1), x0;
initialize m0 ← 0, v0 ← 0, t ← 0
for t ∈ {0,...,maxit} do
    t ← t + 1
    g_t ← ∇f_{i_t}(x_{t-1})
    m_t ← β1 · m_{t-1} + (1 − β1) · g_t
    v_t ← β2 · v_{t-1} + (1 − β2) · g_t^2
    η_t = η √(1−β2^t) / (1−β1^t)
    x_t ← x_{t-1} − η_t · m_t / (√v_t + ε)
end for
Result: x_t
```

Experimental results



Conclusion and perspective work

- Adaptive steplengths make the algorithms more robust then the standard SG methods and provide performances comparable with SG with best-tuned steplengths [6], [5];
- further study to improve the adaptive steplength rules also in the stochastic case;
- validation of the stochastic-Ritz version: experiments on other database and other loss-functions;
- exploit mini-batch of adaptive size; analyse the sensitivity of the step size rules to the mini-batch size, possible combination with inexact Line-Search.

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. Kingma, J. Ba, *Adam: A method for stochastic optimization*, arXiv:1412.6980 (2014)
[4] 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.
[5] Sopyla, Drozda, *SGD with BB update step for SVM*, Inf. Sci., 2015
[6] Tan, Ma, Dai, Qian, *BB Step Size for SGD*, Adv NIPS 2016