Step Size Selection in Frank-Wolfe Method

Paper Review and Replication

Higher School of Economics

Motivation and Problem: Compare different step size selection methods for FW method

This project aims to explore, evaluate and conduct numerical comparison of the efficiency of step selection strategies in the Frank-Wolfe method on different problems with different dimensions and condition numbers. It can solve problems of the form $\min_{x \in Q} f(x),$

We consider four strategies for step size selection and compare corresponding Frank-Wolfe method performance on both real and synthetic data with different properties and constraints on the solution.

As other gradient-based methods, the FW algorithm depends on a step size parameter gamma. The following step size selection approaches were considered:

Proposed methods for compare

- Predefined Decreased (Trivial)

$$\gamma^k = \frac{2}{k+2}$$

- Demyanov-Rubinov

$$\gamma^k = \min\left(\frac{g^k}{L||d^k||^2}, 1\right)$$

- Exact Line-search

$$\gamma^k = \arg\min_{\gamma\in[0,1]} f(x^k + \gamma d^k)$$

- Armijo method

$$f(x^k + \gamma^k d^k) \leq f(x^k) + \sigma\gamma^k \nabla f(x^k)^T d^k$$

Frank-Wolfe Algorithm

Algorithm 1 Frank-Wolfe algorithm

Input: initial guess x_0, gap tolerance $\delta > 0$
for $k = 0, 1, \ldots$ **do**
  $s^k \in \arg\max_{s \in Q} \langle -\nabla f(x^k), s \rangle$
  $d^k = s^k - x^k$
  $g^k = \langle -\nabla f(x^k), d^k \rangle$
  if $g^k < \delta$ **then**
    return x^k
  end if
  Set step size γ^k by the certain selection strategy
  $x^{k+1} = x^k + \gamma^k d^k$
end for
return x^k

Algorithm 2 Frank-Wolfe algorithm with backtracking line-search

Input: initial guess x_0, gap tolerance $\delta > 0$, backtracking line-search parameters $\tau > 1$, $\eta \leq 1$, initial guess for M^{-1}.
for $k = 0, 1, \ldots$ **do**
  $s^k \in \arg\max_{s \in Q} \langle -\nabla f(x^k), s \rangle$
  $d^k = s^k - x^k$
  $g^k = \langle -\nabla f(x^k), d^k \rangle$
  $M^k = \eta M^{k-1}$
  $\gamma^k = \min(\frac{g^k}{M^k ||d^k||^2}, 1)$
  while $f(x^k + \gamma^k d^k) > Q^k(\gamma^k, M^k)$ **do**
    $M^k = \tau M^k$
  end while
  $x^{k+1} = x^k + \gamma^k d^k$
end for
return x^k

$$Q^k(\gamma^k, M^k) = f(x^k) - \gamma^k g^k + \frac{(\gamma^k)^2 M^k}{2} ||d^k||^2, \tag{9}$$

Experiments: Datasets

1. **UCI Mushrooms** (binary classification dataset, contains descriptions of mushrooms (poisonous/edible), 8124 objects, 22 features)
2. **UCI Gisette** (binary classification dataset, contains engineered features of handwritten digits, 13500 objects, 5000 features)
3. **UCI Covertype** (binary classification dataset, contains cartographic variables, 581012 objects, 54 features)
4. **Synthetic normal** (binary classification dataset, generated from scipy package, 5000 objects, 50 features)
5. **Synthetic ill-conditioned** (binary classification dataset, generated from scipy package, 5000 objects, 1024 features)
6. **Synthetic high-dimensional** (binary classification dataset, ill-conditioned, 5000 objects, 50 features)
7. **Rosenbrock function** $f(x) = \sum_{i=1}^{n-1}(100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2)$

Experiments: Setup and constraints

Constraints

For constraint sets on the desired solution we considered l1 and l2 balls centered at 0, of radiuses R = 10, 100, 500

Setup

All the datasets were used without features changes, except instances labels were transformed into 1 and -1 values for positive and negative objects respectively. All the datasets were split on the train and test parts in proportion 8/2. The standard logistic regression objective was considered for the optimization convergence and performance criterion:

$$f(X, Y, w) = \sum_{i=1}^{n} \ell(w^T x_i, y_i), \qquad (11)$$

$$\ell(z, y) = ln(1 + e^{-yz}), \qquad (12)$$

$x_1, \ldots, x_n \in \mathbb{R}^d$ - data objects features, $y_1, \ldots, y_n \in \{-1, 1\}$ - corresponding labels, $w \in \mathbb{R}^d$ - logistic regression models weights.

Mushrooms Dataset Results

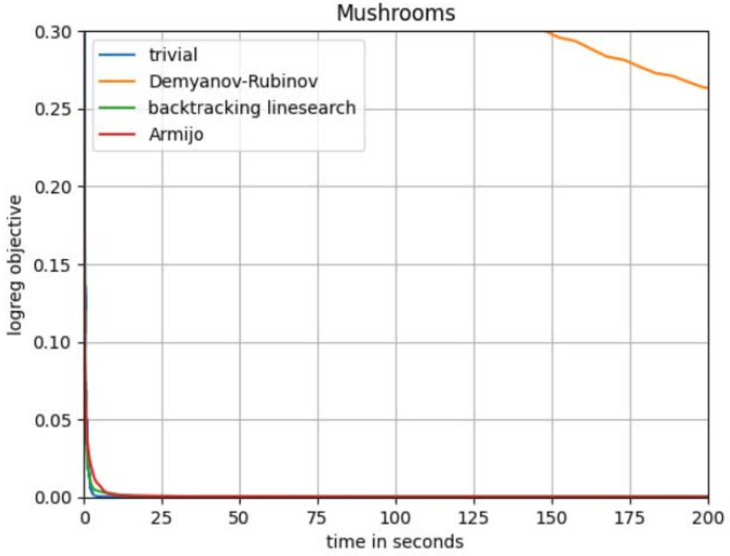

Figure 2. Values of convergence criterion (logreg objective) by time for different Frank-Wolfe method step size values, Mushrooms dataset, constraint on the ℓ_1 ball of radius $R = 100$

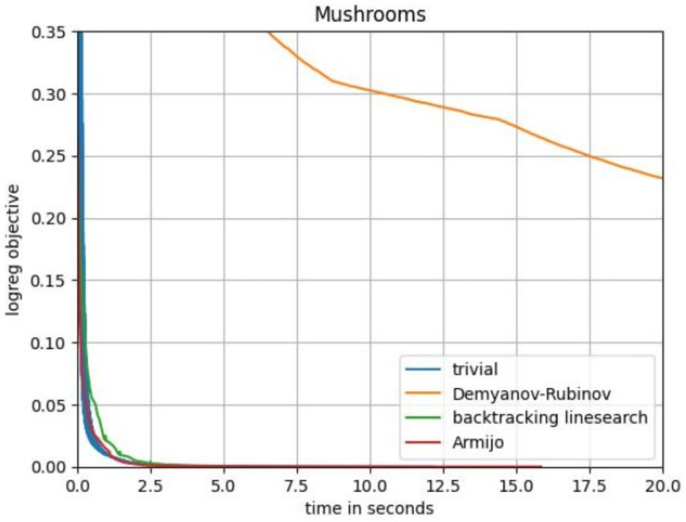

Figure 4. Values of convergence criterion (logreg objective) by time for different Frank-Wolfe method step size values, Mushrooms dataset, constraint on the ℓ_2 ball of radius $R = 100$

Synthetic Normal Dataset Results

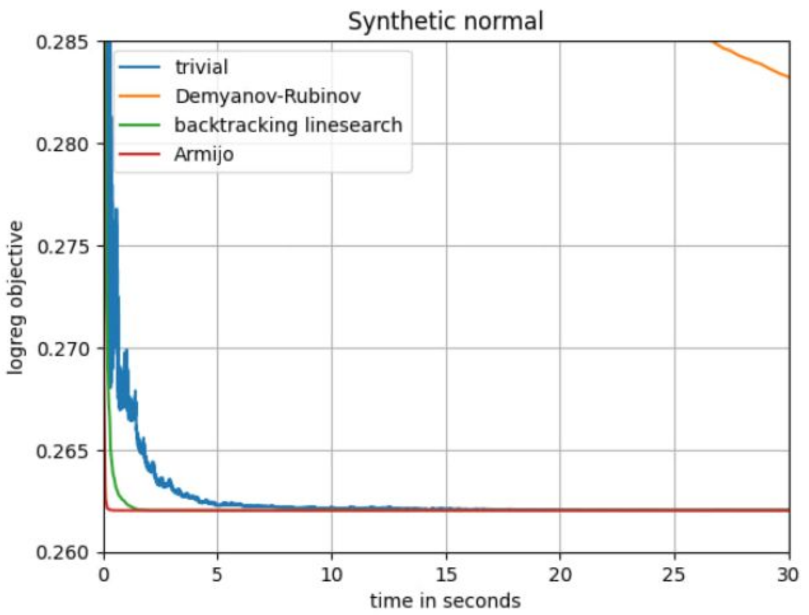

Figure 14. Values of convergence criterion (logreg objective) by time for different Frank-Wolfe method step size values, synthetic normal dataset, constraint on the ℓ_1 ball of radius $R = 100$

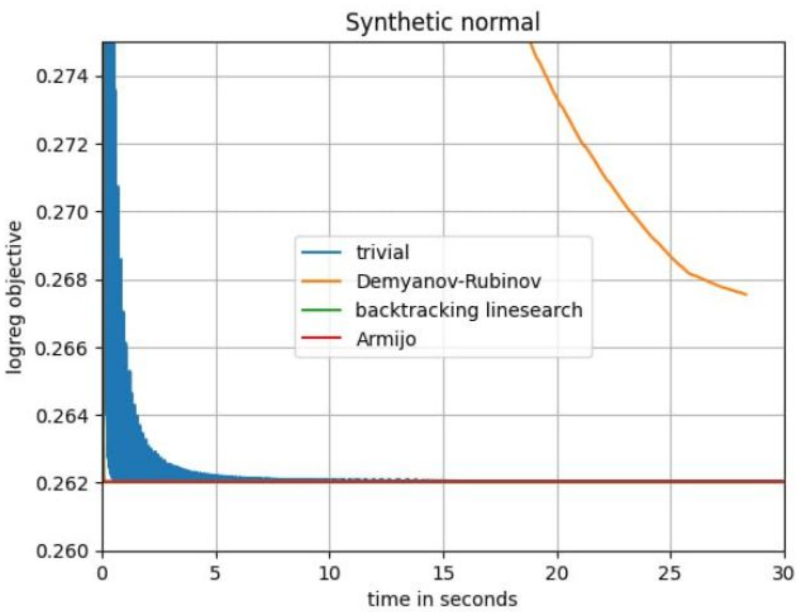

Figure 16. Values of convergence criterion (logreg objective) by time for different Frank-Wolfe method step size values, synthetic normal dataset, constraint on the ℓ_2 ball of radius $R = 100$

Synthetic High-Dimensional Dataset Results

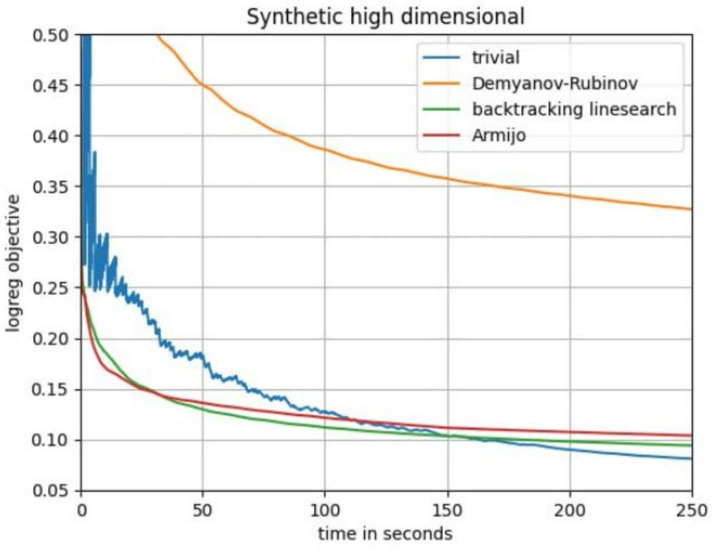

Figure 18. Values of convergence criterion (logreg objective) by time for different Frank-Wolfe method step size values, synthetic high-dimensional dataset, constraint on the ℓ_1 ball of radius $R = 100$

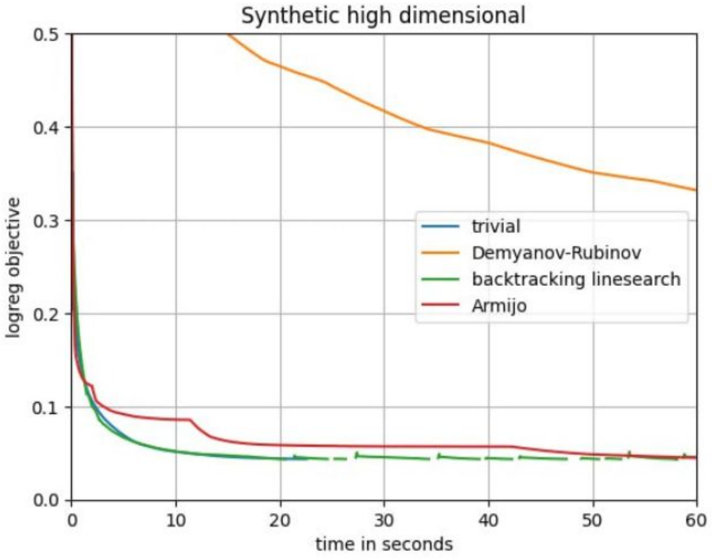

Figure 20. Values of convergence criterion (logreg objective) by time for different Frank-Wolfe method step size values, synthetic high-dimensional dataset, constraint on the ℓ_2 ball of radius $R = 100$

Synthetic ill-Conditioned Dataset Results

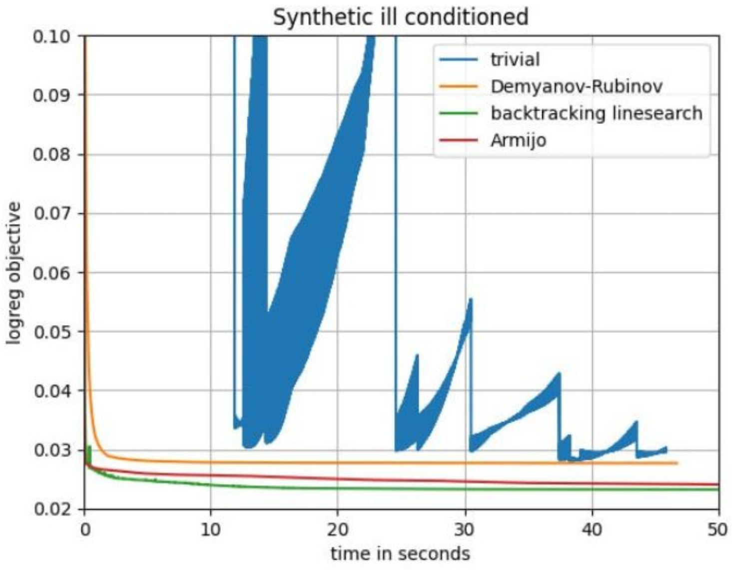

Figure 22. Values of convergence criterion (logreg objective) by time for different Frank-Wolfe method step size values, synthetic ill-conditioned dataset, constraint on the ℓ_1 ball of radius $R = 100$

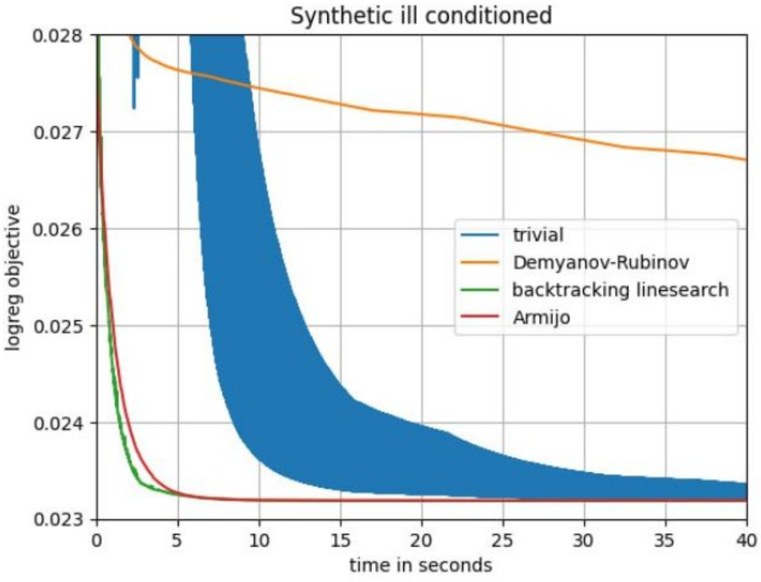

Figure 24. Values of convergence criterion (logreg objective) by time for different Frank-Wolfe method step size values, synthetic ill-conditioned dataset, constraint on the ℓ_2 ball of radius $R = 100$

Gisette + Covtype

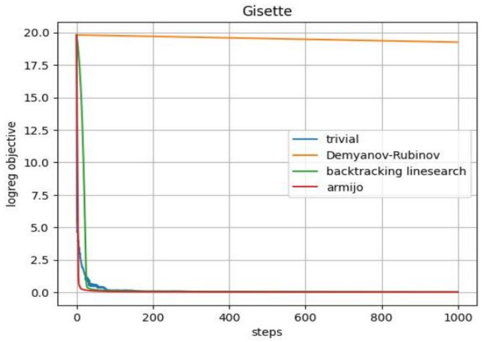

Figure 5. Values of convergence criterion (logreg objective) by iteration number for different Frank-Wolfe method step size values, Gisette dataset, constraint on the ℓ_1 ball of radius $R = 100$

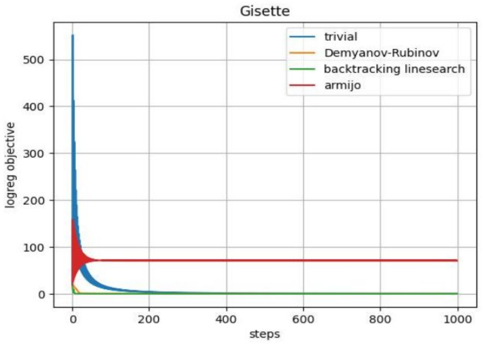

Figure 7. Values of convergence criterion (logreg objective) by iteration number for different Frank-Wolfe method step size values, Gisette dataset, constraint on the ℓ_2 ball of radius $R = 100$

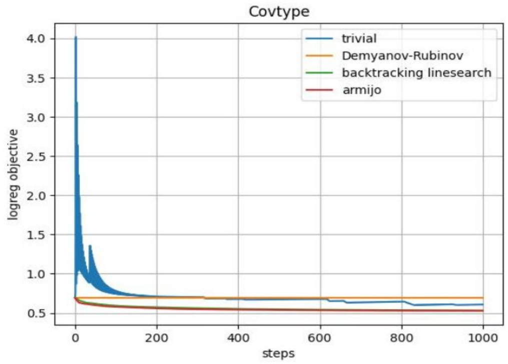

Figure 9. Values of convergence criterion (logreg objective) by iteration number for different Frank-Wolfe method step size values, Covertype dataset, constraint on the ℓ_1 ball of radius $R = 100$

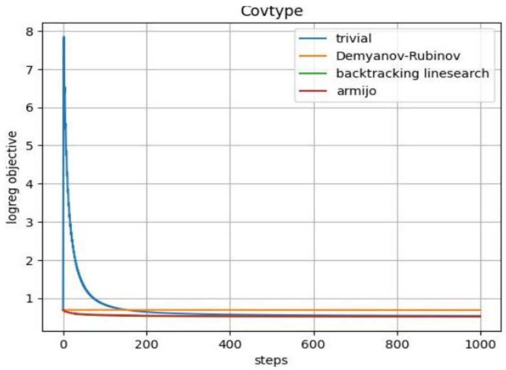

Figure 11. Values of convergence criterion (logreg objective) by iteration number for different Frank-Wolfe method step size values, Covertype dataset, constraint on the ℓ_2 ball of radius $R = 100$

Rosenbrock Results

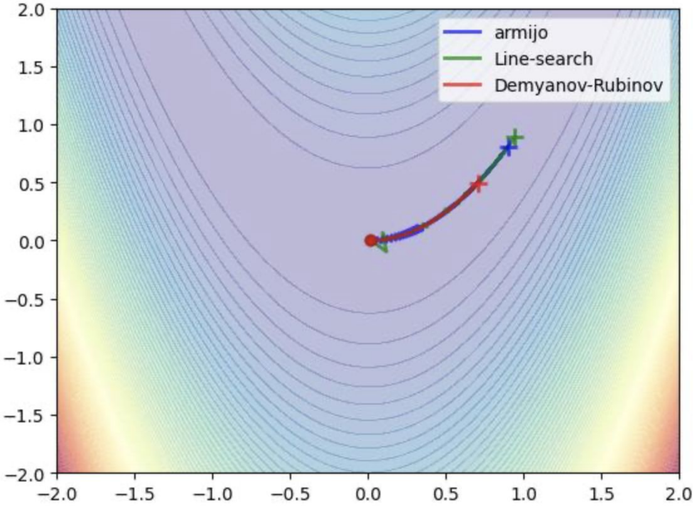

Figure 24. Landscape of convergence for Rosenbrock function for different Frank-Wolfe method step size values, constraint on the l_2 ball of radius R = 100

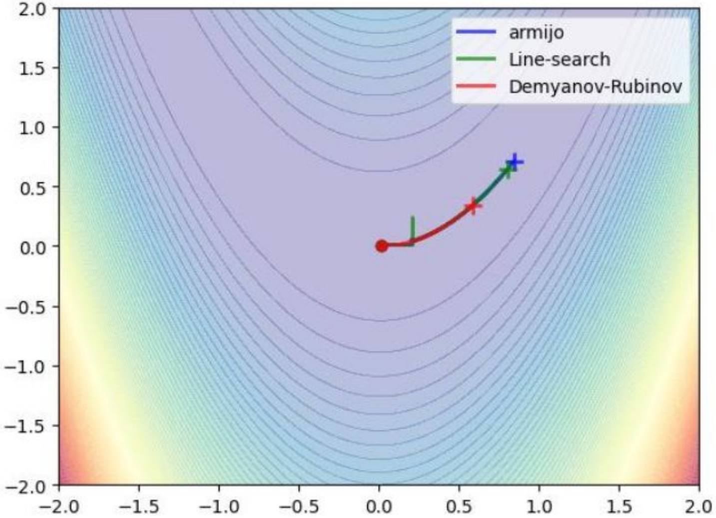

Figure 25. Landscape of convergence for Rosenbrock function for different Frank-Wolfe method step size values, constraint on the ℓ_1 ball of radius $R = 100$

Conclusion

- Backtracking line search has the best performance
- Armijo has the same performance
- Demyanov Rubinov method sometimes even worse that trivial, but on datasets with low L-Lipschitz and low features amount
- High-dimensional dataset leads to unstable convergence
- Functions with complex landscape leads to the bad performance with the trivial approach

Our GitHub: github.com/MarioAuditore/frank_wolfe_step_selection

Our Team

Ignat Romanov
Core algorithm development,
Datasets preparation,
Experiments

Petr Sychev
Theory Research,
Armijo method,
Preso

Boris Miheev
Project Description,
Literature review,
Experiments

Elfat Sabitov
Core algorithm development,
Experiments, GitHub
repository