Stochastic subset conditional gradient

No Author Given

No Institute Given

Abstract. In this paper, we introduce the Stochastic Correspondences Frank-Wolfe method, also known as the Stochastic Subset Conditional Gradient method. This innovative approach addresses the challenge of finding equilibrium flow distributions and extends to a broader range of problems. By reducing the computational complexity of the minimization oracle multiple times, this method trades off increased memory consumption for enhanced computational efficiency. The key advantage lies in minimizing the number of shortest paths that need to be computed. We provide theoretical convergence estimates for the method and demonstrate its practical efficiency through experimental results. Specifically, our findings show that this method significantly outperforms the traditional Frank-Wolfe method and its variants on large datasets. However, it becomes less effective on small datasets, where the cost of finding shortest paths is relatively low. This algorithm proves to be particularly useful for solving transportation problems with large datasets such as the Moscow road network.

Keywords: Convex optimization · Frank-Wolfe algorithm · Conditional gradient

1 Introduction

One of the effective methods for solving the problem of finding equilibrium distribution of traffic flows is the Conditional Gradient method or Frank-Wolfe method. Existing modifications, such as Conjugate Frank-Wolfe [3] or N-Conjugate Frank-Wolfe [1], have a significant advantage over the basic Frank-Wolfe implementation. However, on sufficiently large datasets, the computation time of the linear minimization oracle in the conditional gradient method starts to grow as $O(n \log n)$, where n is the number of edges and vertices of the road graph. Thus the problem of finding more time-efficient methods arises. In this paper, a novel Stochastic Correspondences Frank Wolfe or Stochastic Subset Conditional Gradient method is proposed. The idea behind this method is to make a Frank Wolfe step on a subset of correspondences. This idea is mentioned in [4] in the Randomized Double Averaging Method for Finding Equilibrium in the Stable Dynamics Model.

The key findings of this paper are as follows:

1. The Stochastic Correspondences Frank-Wolfe method is proposed,
2. Two convergence estimates of the method are obtained,
3. Experiments are conducted to show the advantage of the method over Frank-Wolfe.

2 Problem statement

In this paper we consider the problem of finding an equilibrium distribution of flows:

$$\Psi(f) = \sum_{e\in\mathcal{E}} \underbrace{\int_0^{f_e} \tau_e(z)dz}_{\sigma_e(f_e)} \longrightarrow \min_{f=\Theta x,\, x\in X}. \tag{Primal}$$

Although the idea of the proposed algorithm, which will be discussed later, has a more general application, i.e., not only to this problem.

This problem is considered in detail in the article [2]. Let us define some notations. Travel time on roads with a given traffic flow:

$$\tau_e(f_e) = \bar{t}_e\left(1+\rho\left(\frac{f_e}{\overline{f_e}}\right)^{\frac{1}{\mu}}\right) \qquad \tau(f) \equiv \{\tau_e(f_e)\}_{e\in\mathcal{E}}, \tag{Beckman}$$

where $\rho, \mu, \overline{t_e}, \overline{f_e}$ – constants in the Beckman model.

Given a road graph $\mathcal{E}$, the problem is to find the optimal distribution of flows $f^* \equiv \{f_e^*\}_{e\in\mathcal{E}}$.

Assumes that source and destinations pairs and their correspondences are specified:

$$O \text{ and } D - \text{is origin and destination vertices } \subset V \text{ of graph } \mathcal{E}$$
$$OD \equiv \{w \equiv (i,j) | i \in O, j \in D\}$$
$$\mathcal{P} = \{\mathcal{P}_w\}_{w\in OD}.$$

The correspondences of each source- destination pair are additively partitioned into flows along the paths connecting these pairs:

$$X = \{x \mid d_w = \sum_{p\in\mathcal{P}_w} x_p \ ,\ x_p \geqslant 0 \ \forall w \in OD\}.$$

Flows on all paths along some road overlap and give rise to some value of the flow on that road. This can be written compactly by introducing a matrix of edges belonging to paths:

$$\Theta \equiv \{\delta_{e,p}\}_{e\in\mathcal{E}, p\in\mathcal{P}}$$
$$x \equiv \{x_p\}_{p\in\mathcal{P}}$$
$$f = \Theta x,$$

where $\delta_{e,p}$ is a delta function denoting $\mathbf{1}_{e\in p}$, where $e \in \mathcal{E}$ – edge and $p \in \mathcal{P}$ – is a path.

2.1 Frank-Wolfe

Algorithm 1 Frank–Wolfe

```
f_0 ∈ X — starting point
k := 0
repeat
    s_k^FW := argmin_{s∈X} ⟨∇Ψ(f_k), s⟩
    d_k := s_k^FW − f_k
    γ_k := argmin_{γ∈[0,1]} Ψ(f_k + γ · d_k)
    f_{k+1} := f_k + γ_k · d_k
    k := k + 1
until k < max iter
```

In the algorithm, first linear minimization oracle is applied:

$$s_k^{FW} \coloneqq \arg\min_{s\in\mathcal{X}} \langle \nabla\Psi(f_k), s\rangle.$$

Next, the step direction is obtained:

$$d_k = s_k^{FW} - f_k$$

Later, the step length is searched using linesearch and the step is produced:

$$\gamma_k \coloneqq \arg\min_{\gamma\in[0,1]} \Psi(f_k + \gamma\cdot d_k)$$

$$f_{k+1} \coloneqq f_k + \gamma_k \cdot d_k$$

Since $\nabla\Psi(f_k) = \tau(f_k)$, the linear minimization oracle is equivalent to finding shortest paths over all correspondences of OD:

$$\arg\min_{s\in\mathcal{X}} \langle \nabla\Psi(f^k), s\rangle = \arg\min_{s\in\mathcal{X}} \langle \tau(f_k), s\rangle$$

This functional can be decomposed into a sum of independent summands corresponding to individual correspondences. And minimizing it is equivalent to finding the shortest paths, where $\tau(f_k)$ is the cost of traveling along the road:

$$\min_{s\in\mathcal{X}} \langle \tau(f_k), s\rangle = \sum_{w\in OD} T_{p_w^*}(f_k)\cdot d_w,$$

where p_w^* – one of the shortests paths between origin i and destination d of the correspondence $w = (i, j) \in OD$, with current traffic flows f_k on the roads, $T_{p_w^*}(f_k)$ – travel time along the shortest path p_w^*.

Thus for the Frank-Wolfe step it is necessary to find all shortest paths over all correspondences. On large road graphs this can become computationally expensive. The main interest is how to improve the method, to make iteration more computationally efficient.

3 Stochastic subset conditional gradient

Consider optimization problem:

$$f(x) \to \min_{x \in \mathcal{X}}$$

Assumption 1:

$$\mathcal{X} \equiv \{\Theta z | z \in \mathcal{Z}\}$$

For example:

$$\mathcal{X} \equiv \{\Theta x | \sum_{p \in \mathcal{P}_w} x_p = d_w, x_p \geqslant 0\}$$
$$\text{where } \mathcal{Z} = \{x | \sum_{p \in \mathcal{P}_w} x_p = d_w, x_p \geqslant 0\}$$

Assumption 2:

$$\mathcal{X}_v^Q \equiv \{\Theta z, \quad z \in \mathcal{Z} | \forall i \notin Q \to z_i = v_i\}$$
$$y = \arg\min_{s \in \mathcal{X}} \langle r, s \rangle$$
$$y_v^Q = \arg\min_{s \in \mathcal{X}_v^Q} \langle r, s \rangle$$
$$\forall i \in Q \to [\Theta^{-1}(y_v^Q)]_i = [\Theta^{-1}(y)]_i$$

For example:

$$Q = Q_a = \{p | p \in \mathcal{P}_a\}$$
$$\mathcal{Z}_v^Q = \{x \in \mathcal{Z} | x_p = v_p \forall p \notin \mathcal{P}_w\}$$
$$\mathcal{X}_v^Q = \Theta(\mathcal{Z}_v^Q)$$
$$y = \arg\min_{s \in \mathcal{X}} \langle \tau(f), s \rangle \equiv T_{p_a^*}(f) \cdot d_a + \sum_{w \in OD \setminus \{a\}} T_{p_w^*}(f) \cdot d_w$$
$$y_v^Q = \arg\min_{s \in \mathcal{X}_v^Q} \langle \tau(f), s \rangle \equiv T_{p_a^*}(f) \cdot d_a + \sum_{w \in OD \setminus \{a\}} \sum_{p \in \mathcal{P}_w} T_p(f) \cdot s_p$$
$$\forall i \in Q \to [\Theta^{-1}(y_v^Q)]_i = [\Theta^{-1}(y)]_i$$

Assumption 3:

$$y_v^Q = \arg\min_{s \in \mathcal{X}_v^Q} \langle r, s \rangle$$
$$y = \arg\min_{s \in \mathcal{X}} \langle r, s \rangle$$
$$\mathbb{E} y_v^Q = \frac{m}{n} y + (1 - \frac{m}{n}) v, \quad \text{where } m = |Q|$$

So, the Frank-Wolfe method for computing the minimization oracle requires to calculate the shortest paths for all correspondences. Since on large road graphs the search for shortest paths takes large computational resources, we need to obtain an algorithm that uses only a fraction of all correspondences. Note that Dijkstra's algorithm finds the shortest paths from a given vertex to all other vertices. Therefore, it makes sense to choose not just a random correspondence, but to choose a random source and count the shortest paths through all outgoing correspondences. In this way the computational cost is reduced in the number of sources. Further, the experiments will randomly select a certain fraction from the set of sources.

Using the assumptions above we can formulate the Stochastic Correspondences Frank-Wolfe method:

Algorithm 2 Stochastic Correspondences Frank–Wolfe

$x_0 \in \mathcal{X}$ — starting point.
$k := 0$
repeat
  $t_1, ..., t_m \sim U[1, n] \quad t_i \neq t_j$
  $Q = \{p_{t_1}, ..., p_{t_m}\} \subset \mathcal{P}$
  $s_k^{FW} := \arg\min\limits_{s \in \mathcal{X}_{x_k}^Q} \langle \nabla f(x_k), s \rangle$
  $d_k^Q := s_k^{FW} - x_k^Q$
  $\gamma_k := \arg\min\limits_{\gamma \in [0,1]} \Psi(x_k^{\bar{Q}} + x_k^Q + \gamma \cdot d_k^Q)$
  $x_{k+1}^Q := x_k^Q + \gamma_k \cdot d_k^Q$
  $k := k + 1$
until $k <$ max iter

4 Method convergence

Theorem 1. *Let the optimization problem satisfies assumption 1 and one of assumptions 2 or 3. And let $f(x_0) - f(x^*) \leqslant \frac{LR^2}{\alpha^2}$. Then this method converges such as:*

$$\mathbb{E} f(x^N) - f(x^*) \leqslant \frac{2LR^2}{N+2} \cdot \frac{1}{\alpha^2}$$

Proof.

$$f(x_{k+1}) \leqslant f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|_2^2$$

$$f(x_{k+1}) \leqslant f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|_2^2$$

$$\mathbb{E}[f(x_{k+1}) | x_k] \leqslant f(x_k) + \gamma \langle \nabla f(x_k), \mathbb{E} s_k - x_k \rangle + \frac{\gamma^2 LR^2}{2}$$

Use Assumption 3 or consequence Assumption 2, and $\alpha = \frac{|Q|}{n} = \frac{m}{n}$:

$$\mathbb{E}[f(x_{k+1})|x_k] \leqslant f(x_k) + \gamma\langle\nabla f(x_k), \alpha s_k^{FW} + (1-\alpha)x_k - x_k\rangle + \frac{\gamma^2 LR^2}{2}$$

$$\mathbb{E}[f(x_{k+1})|x_k] \leqslant f(x_k) + \gamma\alpha\langle\nabla f(x_k), s_k^{FW} - x_k\rangle + \frac{\gamma^2 LR^2}{2}$$

Use definition of s_k^{FW} and convexity of function:

$$\langle\nabla f(x_k), s_k^{FW} - x_k\rangle \leqslant \langle\nabla f(x_k), x^* - x_k\rangle \leqslant f(x^*) - f(x_k)$$

$$\mathbb{E}[f(x_{k+1})|x_k] \leqslant f(x_k) + \gamma\alpha(f(x^*) - f(x_k)) + \frac{\gamma^2 LR^2}{2}$$

Let apply expectaton by x_k:

$$\mathbb{E}f(x_{k+1}) = \mathbb{E}[\mathbb{E}[f(x_{k+1})|x_k]] \leqslant \mathbb{E}f(x_k) + \gamma\alpha(f(x^*) - \mathbb{E}f(x_k)) + \frac{\gamma^2 LR^2}{2}$$

$$\mathbb{E}f(x_{k+1}) - f(x^*) = (1-\gamma\alpha)(\mathbb{E}f(x_k) - f(x^*)) + \frac{\gamma^2 LR^2}{2}$$

Lets define $\delta_k = \frac{\mathbb{E}f(x_k) - f(x^*)}{LR^2}$:

$$\delta_{k+1} \leqslant (1-\gamma\alpha)\delta_k + \frac{\gamma^2}{2}$$

Use assumption $\delta_0 \leqslant \frac{1}{\alpha^2}$ and move on continious case $\delta(t)$:

$$\delta_{k+1} - \delta_k \leqslant \min_{\gamma\in[0,1]}[\frac{\gamma^2}{2} - \gamma\alpha\delta_k]$$

$$\delta_{k+1} - \delta_k \leqslant -\frac{\alpha^2\delta_k^2}{2}$$

$$\delta(t+1) - \delta(t) \leqslant -\frac{\alpha^2\delta(t)^2}{2}$$

Use convexity of $\delta(t)$:

$$\delta(t)' \leqslant \delta(t+1) - \delta(t) \leqslant -\frac{\alpha^2\delta(t)^2}{2}$$

Solve differential equation:

$$\frac{\alpha^2 t}{2} \leqslant -\int_0^t \frac{\delta(t)'}{\delta(t)^2}dt = \frac{1}{\delta(t)} - \frac{1}{\delta(0)}$$

$$\delta(t) \leqslant \frac{2}{t + \frac{2}{\alpha^2\delta(0)}} \cdot \frac{1}{\alpha^2}$$

Consequently:

$$\mathbb{E}f(x_N) - f(x^*) = LR^2 \cdot \delta_N \leqslant \frac{2LR^2}{N + \frac{2}{\alpha^2\delta(0)}} \cdot \frac{1}{\alpha^2}$$

$$\alpha^2\delta(0) \leqslant 1$$

$$N + \frac{2}{\alpha^2\delta(0)} \geqslant N + 2$$

$$\mathbb{E}f(x_N) - f(x^*) \leqslant \frac{2LR^2}{N+2} \cdot \frac{1}{\alpha^2}$$

For example in the case when $\alpha = 1$, we can get classic convergence of Frank-Wolfe:

$$\text{Stochastics dissapear: } f(x_N) - f(x^*) = \mathbb{E}f(x_N) - f(x^*)$$

$$f(x_N) - f(x^*) \leqslant \frac{2LR^2}{N+2}$$

□

Theorem 2. *This method also converges such as :*

$$\min_{k \in 1,\dots,N} \mathbb{E}g_k \leqslant \frac{\min\{2h_0, LR^2\}}{\sqrt{N+1}} \cdot \frac{1}{\alpha},$$

where $g_k = -\min_{s \in \mathcal{X}} \langle \nabla f(x_k), s - x_k \rangle$ *– Frank-Wolfe gap.*

Proof. Let use the formula from the previous proof:

$$\mathbb{E}[f(x_{k+1})|x_k] \leqslant f(x_k) + \gamma\alpha\langle \nabla f(x_k), s_k^{FW} - x_k \rangle + \frac{\gamma^2 LR^2}{2}$$

$$g_k = -\langle \nabla f(x_k), s_k^{FW} - x_k \rangle$$

$$\mathbb{E}f(x_{k+1}) \leqslant \mathbb{E}f(x_k) - \gamma\alpha\mathbb{E}g_k + \frac{\gamma^2 LR^2}{2}$$

$$\mathbb{E}f(x_{k+1}) \leqslant \mathbb{E}f(x_k + \gamma d_k) \leqslant \mathbb{E}f(x_k) - \gamma\alpha\mathbb{E}g_k + \frac{\gamma^2 LR^2}{2} \quad \forall \gamma \in [0,1]$$

$$\mathbb{E}f(x_{k+1}) \leqslant \mathbb{E}f(x_k) + \min_{\gamma \in [0,1]} \left\{ -\gamma\alpha\mathbb{E}g_k + \frac{\gamma^2 LR^2}{2} \right\}.$$

$$G_k \equiv \mathbb{E}g_k$$

$$\gamma^* = \frac{\alpha G_k}{LR^2} \to \mathbb{E}f(x_{k+1}) \leqslant \mathbb{E}f(x_k) - \frac{\alpha^2 G_k^2}{2A}, \qquad \text{if } \alpha G_k \leqslant LR^2$$

$$\gamma^* = 1 \to \mathbb{E}f(x_{k+1}) \leqslant \mathbb{E}f(x_k) - \frac{\alpha G_k}{2}, \qquad \text{if } \alpha G_k > LR^2$$

$$
\begin{aligned}
G_N^* &\equiv \min_{k \in \overline{1,\dots,N}} G_k \\
-h_0 \leqslant \mathbb{E} f(x_N) - f(x_0) &\leqslant -\sum_{k=0}^{N} \frac{\alpha G_k}{2} \min\left\{\frac{\alpha G_k}{LR^2}, 1\right\} \\
&\leqslant -(N+1) \cdot \frac{\alpha G_N^*}{2} \min\left\{\frac{\alpha G_N^*}{LR^2}, 1\right\}.
\end{aligned}
$$

$$
\begin{cases}
\alpha G_N^* \leqslant \frac{\sqrt{2h_0 LR^2}}{\sqrt{N+1}} \leqslant \frac{2h_0 + LR^2}{2\sqrt{N+1}}, & \text{if } \alpha G_N^* \leqslant LR^2, \\
\alpha G_N^* \leqslant \frac{2h_0}{N+1} \leqslant \frac{2h_0}{\sqrt{N+1}}, & \text{if } \alpha G_N^* > LR^2.
\end{cases} \tag{1}
$$

Taking the maximum of rhs gives (2).

□

5 Experiments

Experiments were conducted on large datasets from the TransportationNetworks repository: Chicago-Regional, GoldCoast, Birmingam-England, Chicago-Sketch, and Philadelphia.

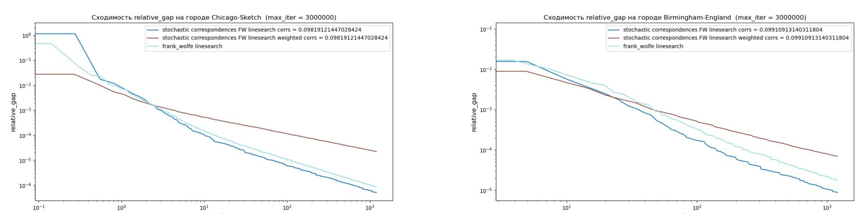

Fig. 1. Caption

Experiments show that:

1. SCFW has an advantage on large road graphs over Frank-Wolfe ,
2. It follows that, in practice, it is more efficient to update the pitch frequently, but only on a subset of correspondences, than to update the pitch infrequently, but compute the honest Frank-Wolfe minmization oracle.
3. The quality gap between SCFW and FW grows as the size of the road graph increases.
4. SCFW loses on small graphs.

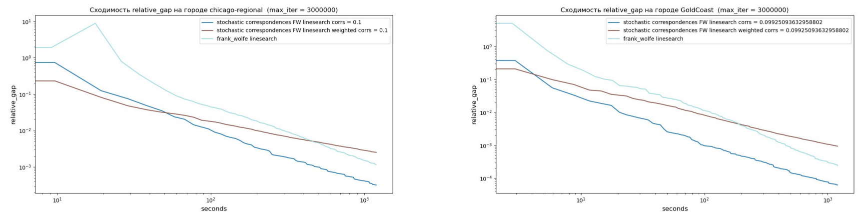

Fig. 2. Caption

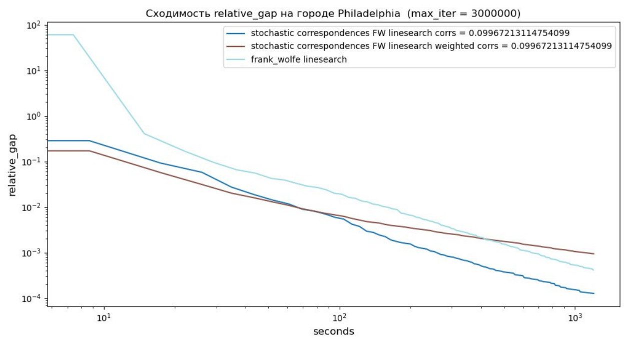

Fig. 3. Caption

6 Conclusion

References

1. Ignashin, I., Yarmoshik, D.: Modifications of the frank-wolfe algorithm in the problem of finding the equilibrium distribution of traffic flows. Mathematical Modeling and Numerical Simulation **10**(1), 10–25 (2024)
2. Kotlyarova E.V., Gasnikov A.V., G.E.Y.D.: Finding equilibria in two-stage models of traffic flow distribution over a network. In: Computer Research and Modeling. pp. vol. 28, no. 10, 40–64 (2021)
3. Mitradjieva, M., Lindberg, P.O.: The stiff is moving—conjugate direction frank-wolfe methods with applications to traffic assignment. Transportation Science **47**(2), 280–293 (2013)
4. . V. Gasnikov, P. E. Dvurechenskii, Y.V.D.Y.V.M.: Numerical methods of searching for equilibrium distribution of flows in the beckman model and in the model of stable dynamics. In: Mathematical Modeling. pp. vol. 28, no. 10, 40–64 (2016)