Boosting Frank-Wolfe by Chasing Gradients
论文实现方法讨论及案例复现

高世杰，晏浩洋

keyword:Frank-Wolfe,boosting,align,convergence analysis

Abstract

本文旨在分析《Boosting Frank-Wolfe by Chasing Gradients》一文中对 Frank-Wolfe 算法进行了怎样的改进。文章内容包括 [1]Frank-Wolfe 算法存在的问题及改进思路 [2] 具体算法的各步骤的推导及思想 [3] 算法理论收敛性和实际观测收敛性证明推导 [4] 论文案例复现 [5] 算法意义及可改进方向五个部分。

1. Background and Motivation

1.1. 带有凸约束的凸问题

对于无约束凸问题，最优解 x^* 等价于 $\nabla f(x^*)=0$，基于该定理，产生了各种不同算法，其中一阶算法中较为广泛应用的有梯度下降法。对于带有凸约束的凸问题，无法直接使用梯度下降法，但是梯度下降的想法是可借鉴的，因此产生了两种类型的算法。第一种是梯度投影法，通过计算梯度并将其投影回可行域，此方法每步迭代下降量大，但是有着投影计算量大的问题。第二种方法是 Frank-wolfe 算法，通过选择可行域内与负梯度内积最大的方向进行迭代，此方法的好处在于只需起始点在可行域内，此后的迭代点均在可行域内，不需要投影即可满足解可行，单步计算量小，但是每步迭代方向无法保证充分的下降。

1.2. Frank-Wolfe 算法

Let $\mathcal{H}$ be a Euclidean space,consider

$$\min f(x)$$

Algorithm 1 Frank-Wolfe(FW)

Input: Start point $x_0 \in \mathcal{C}$, $\gamma_t \in [0,1]$.

for $i=1$ to ... do

$$v_t \leftarrow \arg\max_{v\in\mathcal{V}} \langle \nabla f(x_t), v\rangle \quad (1)$$

$$x_{t+1} \leftarrow x_t + \gamma_t (v_t - x_t) \quad (2)$$

end for

$$\text{s.t. } x \in \mathcal{C}$$

$f: \mathcal{H} \to R$ is a smooth convex function

$\mathcal{C} \subset \mathcal{H}$ is a compact convex set, $\mathcal{C} = \text{conv}(\mathcal{V})$

对于多面体约束而言,(1) 式中 $v \in \mathcal{C}$ 可写成 $v \in \mathcal{V}$，因为 v 总在顶点处取得；对于非多面体而言，令 $\mathcal{V} := \partial\mathcal{C}$(边界), 则有同样结果。

(1) 意为寻找迭代方向，满足:

$v_t \leftarrow \arg\max_{v\in\mathcal{C}} \langle \nabla f(x_t), v-x_t\rangle \Leftrightarrow \arg\max_{v\in\mathcal{V}} \langle \nabla f(x_t), v\rangle$

d_t=$v_t - x_t$ 为迭代方向

(2) 对 x_{t+1} 进行迭代，有性质:

$$\begin{aligned} If\ x_t \in \mathcal{C}, x_{t+1} &= x_t + \gamma_t d_t \\ &= x_t t + \gamma_t (v_t - x_t) \\ &= (1-\gamma_t) x_t + \gamma_t v_t \in \mathcal{C} \end{aligned}$$

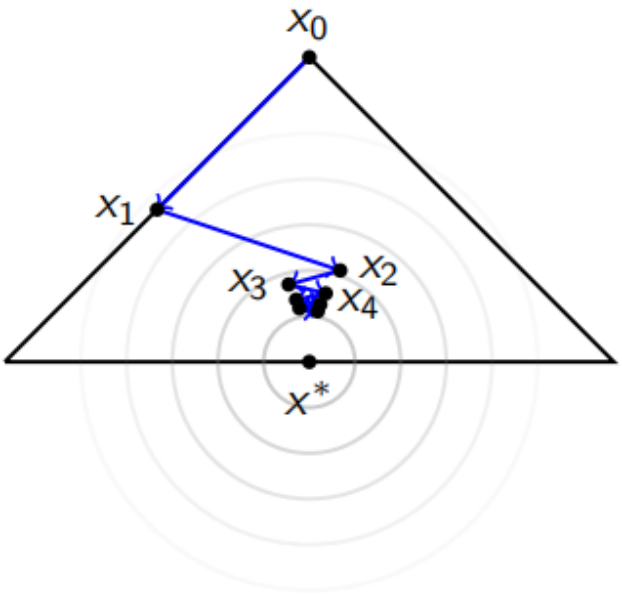
1.3. Frank-Wolfe 算法的不足与改进切入点

Frank-Wolfe 有着单步计算量小 (不计算梯度，仅使用已有信息包括约束集顶点 $\mathcal{V}$，当前点梯度值 $\nabla f(x_t)$) 的优点，但是却无法保证每步迭代充分下降。考虑以下问题:

$$\min \tfrac{1}{2}\|x\|_2^2$$

$$\text{s.t.} x \in \text{conv}\left(\begin{pmatrix}0\\1\end{pmatrix}, \begin{pmatrix}-1\\0\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}\right)$$

start point:$x_0 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$



由于无法保证移动方向与负梯度方向拟合较好，很可能会出现 zig-zagging 的情况。如果能够较好拟合负梯度方向，就能够基于梯度下降法的保证，获得较好的收敛性。因此如何在只使用与 Frank-Wolfe 相同的信息 $(\mathcal{V}, \nabla f(x_t))$ 的情况下较好的拟合负梯度方向便是直观上可行的一种改进措施，也正是《Boosting Frank-Wolfe by Chasing Gradients》一文的切入点。

2. Boosting Frank-Wolfe **算法**

2.1. **符号与定理**

(i)$D := \max_{x,y\in\mathcal{C}} \|y - x\|$

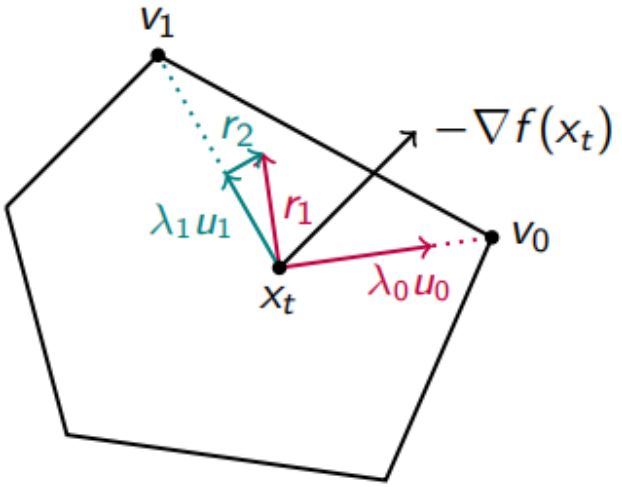
(ii)L-smooth if $L > 0$ and for all $x, y \in \mathcal{H}$,
$$f(y) - f(x) - \langle \nabla f(x), y - x\rangle \leq \frac{L}{2}\|y - x\|^2$$

(iii)S-strongly convex if $S > 0$ and for all $x, y \in \mathcal{H}$,
$$f(y) - f(x) - \langle \nabla f(x), y - x\rangle \geq \frac{S}{2}\|y - x\|^2$$

(iv)μ-gradient dominated if $\mu > 0$,and for all $x \in \mathcal{H}$,
$$f(x) - \min_{\mathcal{H}} f \leq \frac{\|\nabla f(x)\|^2}{2\mu}$$

2.2. **直观算法理解**

Boosting Frank-Wolfe 算法针对 Frank-Wolfe 算法中的 (1) 即迭代方向进行了修正，通过不断将负梯度及负梯度分解后的残差投影至顶点与当前点连线方向，既只是用了 $(\mathcal{V}, \nabla f(x_t), x_t)$ 的信息，又通过不断缩减残差的方式使得迭代方向能够逼近负梯度方向，最后通过一定的参数调整，保证下一步迭代点在可行域中。考虑以下问题:



(1) $v_0 \leftarrow \arg\max_{v\in\mathcal{V}} \langle \nabla f(x_t), v\rangle$

(2) $\lambda_0 u_0 = \frac{\langle -\nabla f(x_t), v_0 - x_t\rangle}{\|v_0 - x_t\|^2}(v_0 - x_t)$

(3) $r_1 = -\nabla f(x_t) - \lambda_0 u_0$

(4) $v_1 \leftarrow \arg\max_{v\in\mathcal{V}} \langle r_1, v\rangle$

(5) $\lambda_1 u_1 = \frac{\langle r_1, v_1 - x_t\rangle}{\|v_1 - x_t\|^2}(v_1 - x_t)$

(6) $r_2 = r_1 - \lambda_1 u_1$

We could continue count:v_2

1. 选择与负梯度相乘最大的顶点 v_0, 将 $v_0 - x_t$ 作为初始拟合负梯度方向

2. 对负梯度在该方向上进行投影，记 $u_0 = v_0 - x_t$，投影系数可以直接通过投影公式计算得到，记 $\lambda_0 = \text{proj}_{u_0}(-\nabla f(x_t)) = \frac{\langle -\nabla f(x_t), v_0 - x_t\rangle}{\|V_0 - x_t\|^2}$

3. 记拟合残差为 r_1

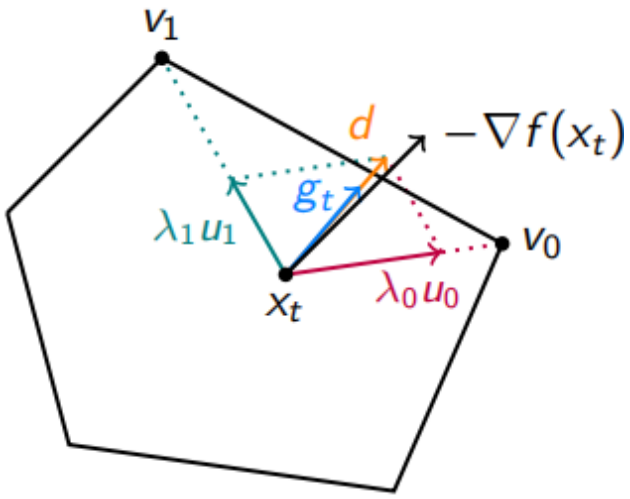
4. 选择与残差 r_1 相乘最大的顶点 v_1, 将 $v_1 - x_t$ 作

为拟合 r_1 方向

5. 对残差进行投影，并计算投影系数

6. 记 r_1 的拟合残差为 r_2

可以看到在拟合了两轮后，拟合方 $d=\lambda_0 u_0+\lambda_1 u_1$ 与 $-\nabla f(x_t)$ 仅仅相差 r_2，比最初的 r_1 已经小了很多，迭代方向的选择得到了明显的改善



(7) $d=\lambda_0 u_0+\lambda_1 u_1$

(8) $g_t=d/(\lambda_0+\lambda_1)$

若不再进行残差拟合，选择 $d=\lambda_0 u_0+\lambda_1 u_1$ 作为最终迭代方向，则 $g_t=d/(\lambda_0+\lambda_1)$ 作为迭代步，可保证 $x_{t+1}\in\mathcal{C}$，详细证明在稍后给出。

2.3. Boosting Frank-Wolfe **算法**

(1) 初始化 $x_0\in\mathcal{C}$

对一次迭代:

(2) 初始化负梯度拟合方向 $d_0=0$。初始放缩系数 $\Lambda_t=0$，用于最终调整 $g_t=d/\Lambda_t$ 满足 $x_t+g_t\in\mathcal{C}$。flag 设立用于判断程序是提前退出还是执行完全，用于确定迭代次数

K 是人为规定的最多次拟合迭代方向次数

(3) 初始化当前轮需要拟合的残差，特别的当 $k=0$ 时，残差 $r_0=-\nabla f(x_t)$

(4) 选择与残差 r_k 相乘最大的顶点 v_k, 将 v_k-x_t 作为拟合 r_k 方向

Algorithm 2 Boosted Frank-Wolfe (BoostFW)

Input: Input point $y\in\mathcal{C}$, maximum number of rounds $K\in N^*$, alignment improvement tolerance $\delta\in[0,1]$, step-size strategy $\gamma_t\in[0,1]$.

$x_0\leftarrow\arg\min_{v\in\mathcal{V}}\langle\nabla f(y),v\rangle$ (1)

for $t=0$ to $T-1$ do

$d_0\leftarrow 0\quad\Lambda_t\leftarrow 0\quad$ flag $\leftarrow$ false (2)

for $k=0$ to $K-1$ do

$r_k\leftarrow-\nabla f(x_t)-d_k$ (3)

$v_k\leftarrow\arg\max_{v\in\mathcal{V}}\langle r_k,v\rangle$ (4)

$u_k\leftarrow\arg\max_{u\in\{v_k-x_t,-d_k/\|d_k\|\}}\langle r_k,u\rangle$ (5)

$\lambda_k\leftarrow\frac{\langle r_k,u_k\rangle}{\|u_k\|^2}$ (6)

$d_k'\leftarrow d_k+\lambda_k u_k$ (7)

if $\text{align}(-\nabla f(x_t),d_k')-\text{align}(-\nabla f(x_t),d_k)\geq\delta$ then

$d_{k+1}\leftarrow d_k'$ (8)

$\Lambda_t\leftarrow\begin{cases}\Lambda_t+\lambda_k & \text{if } u_k=v_k-x_t\\ \Lambda_t(1-\lambda_k/\|d_k\|) & \text{if } u_k=-d_k/\|d_k\|\end{cases}$ (9)

else

flag $\leftarrow$ true break

end if

end for

$K_t\leftarrow k$ if flag = true else K (10)

$g_t\leftarrow d_{K_t}/\Lambda_t$ (11)

$x_{t+1}\leftarrow x_t+\gamma_t g_t$ (12)

end for

(5) $u_k=v_k-x_t$ 为当前选择残差拟合方向

(6) λ_k 为 r_k 在 u_k 上的投影系数

(7) 记 $d_k'=d_k+\lambda_k u_k$ 为待定更新拟合方向

$\text{align}(d,\hat{d}):=\begin{cases}\frac{\langle d,\hat{d}\rangle}{\|d\|\|\hat{d}\|} & \text{if } \hat{d}\neq 0\\ -1 & \text{if } \hat{d}=0\end{cases}$ $\text{align}(d,\hat{d})$ 代表了 $(d,\hat{d})$ 归一化后的相近程度，也即两个向量的夹角的余弦值 $\in[-1,1]$，结合算法 IF 条件来看。IF 条件意为规定每次拟合都要至少进行一定程度的改进 δ，若没有达到预设值，则认为本次包括之后可能的拟合都不再有有效的改进，退出 x_t 点处迭代方向拟合，选择最后拟合方向作为迭代方向。align 定

义中的-1 保证了方向拟合至少进行一轮，即 $d_0 = 0$ 时 $\text{align}(-\nabla f, d_0) = -1$ 保证第一轮一定更新。此设定目的为防止 $d = 0, \lambda = 0$。且由于人为设定 K 值及更新条件，因此最多进行 $\min(\text{K}, \lceil \frac{1}{\delta} \rceil)$ 次拟合。继续来看若改进值满足条件:

(8) 将待定更新方向更改为当前负梯度拟合方向

(9) 更新放缩系数 $\Lambda_{t+1} = \Lambda_t + \lambda_k$ 11 中的备选方案以及 (7) 中的 $-d_k / \|d_k\|$ 目的均为保证算法的收敛的特殊取值情况，也保证了 $\lambda_k \neq 0$

(10) 确定更新次数，若算法提早结束，则负梯度拟合向量为 d_{K_t}，否则为 d_{K_t+1}，$K_t = K - 1$

(11) 将拟合负梯度方向进行放缩使其满足 $x_t + g_t = d_{K_t} / \Lambda_t \in \mathcal{C}$

(12) 选择步长对迭代点进行更新，其中 $x_{t+1} \in \mathcal{C}$

特别的当拟合阶段仅拟合一次，即取 $d = d_1$ 时算法退化为 Frank-Wolfe 算法

2.4. 算法性质

Lemma: $x^* = \arg\min_{v \in \mathcal{V}} \langle r, v \rangle = \arg\min_{z \in \mathcal{C}} \langle r, v \rangle, \forall r \in \mathcal{R}^n$.Argmax has the same result.

$$\forall z \in C, C = \text{conv}(\mathcal{V})$$

$$z = \sum_{v_i \in \mathcal{V}} a_{v_i} v_i, \sum_{v_i \in \mathcal{V}} a_{v_i} = 1$$

$$\begin{aligned} \langle r, x^* \rangle - \langle r, z \rangle &= \langle r, x^* \rangle - \langle r, \sum_{v_i \in \mathcal{V}} a_{v_i} v_i \rangle \\ &\leq \langle r, x^* \rangle - \langle r, \sum_{v_i \in \mathcal{V}} a_{v_i} v_i \rangle \\ &= 0 \end{aligned}$$

$$x^* = \arg\min_{v \in \mathcal{V}} \langle r, v \rangle = \arg\min_{z \in \mathcal{C}} \langle r, v \rangle, \forall r \in \mathcal{R}^n$$

Theorem1 $\forall\, k \in N, \langle r_k, u_k \rangle \geq 0, \lambda_k > 0$

By algorithm, $v_k = \arg\max_{v \in \mathcal{V}} \langle r_k, v \rangle = \arg\max_{z \in \mathcal{C}} \langle r_k, z \rangle$

$$\begin{aligned} \langle r_k, u_k \rangle &= \langle r_k, v_k \rangle - \langle r_k, x_t \rangle \\ &\geq 0 \ (By\ Lemma) \end{aligned}$$

$$\lambda_k = \frac{\langle r_k, u_k \rangle}{\|u_k\|^2} > 0$$

Since if $\lambda_k = 0$ *then* $r_k = 0$ *or* $\langle r_k, u_k \rangle = 0$, *the algorithm is already convergence by algorithm.*

Theorem2 $If\ x_t \in \mathcal{C}, x_t + g_t \in \mathcal{C}$

Define K_t *as times of update of* d, *by algorithm, we have* $K_t \geq 1$

$$\begin{aligned} d &= \sum_{k=0}^{k_t - 1} \lambda_k (v_k - x_t) \\ g_t &= \frac{1}{\Lambda_t} \sum_{k=0}^{k_t - 1} \lambda_k (v_k - x_t) \\ &= \frac{1}{\Lambda_t} \sum_{k=0}^{k_t - 1} \lambda_k v_k - x_t \end{aligned}$$

$$\frac{1}{\Lambda_t} \sum_{k=0}^{k_t - 1} \lambda_k v_k \in \mathcal{C}\ for\ v_k \in \mathcal{C}\ and\ \frac{1}{\Lambda_t} \sum_{k=0}^{k_t - 1} \lambda_k = 1$$

$$x_t + g_t \in \mathcal{C}$$

Theorem3 $If\ x_t \in \mathcal{C}, x_{t+1} \in \mathcal{C}$

$$Let\ y_t = \frac{1}{\Lambda_t} \sum_{k=0}^{k_t - 1} \lambda_k v_k$$

$$\begin{aligned} x_{t+1} &= x_t + \gamma_t g_t \\ &= x_t t + \gamma_t (y_t - x_t) \\ &= (1 - \gamma_t) x_t + \gamma_t y_t \in \mathcal{C} \end{aligned}$$

2.5. 算法小结

Boosting Frank-Wolfe 算法对 Frank-Wolfe 算法的 (1) 进行了改进，使迭代方向的选择更加贴近负梯度方向，保证了每步迭代的下降量，同时没有增加太多的计算量，所使用的信息 $(\mathcal{V}, \nabla f(x_t), x_t)$ 均为已有且与 Frank-Wolfe 算法中使用的相同信息，因此在实验阶段可以看到 Boosting Frank-Wolfe 算法在迭代次数和 CPU 运算时间上均有良好表现。

3. 收敛性分析

3.1. 收敛性结果

传统的 Frank-Wolfe 算法在 convex and L-smooth 条件下为 $O(\frac{1}{\varepsilon})$ 即次线性收敛。在 convex, L-smooth 和 μ-gradient 条件下，Boosting Frank-Wolfe 算法最差情况与 Frank-Wolfe 算法收敛速率相同，即次线性收敛。但实际实验观测结果为 $O\left(\ln \frac{1}{\varepsilon}\right)$ 线性收

敛，实现此收敛速率需要加上一个条件：拟合负梯度方向超过一次的迭代点在所有迭代点中占到一定数量 (也可以说是使用了加速程序的迭代点，因为当拟合次数 K=1 时 Boosting Frank-Wolfe 算法退化为 Frank-Wolfe 算法)，这也符合实验观测结果。

3.2. 引理

3.2.1. 引理及其推导

Theorem4 Let $f : \mathcal{H} \to R$ be L-smooth, convex, and μ-gradient dominated, and set $\gamma_t \leftarrow min\{\eta_t\|\nabla f(x_t)\|/(L\|g_t\|), 1\}$ or $\gamma_t \leftarrow \arg\max_{\gamma\in[0,1]} f(x_t + \gamma g_t)$.Then $\forall$ t $\in [0, T]$

$$f(x_t) - \min_{\mathcal{C}} f \leq \frac{LD^2}{2}\prod_{s=0}^{t-1}\left(1-\eta_s^2\frac{\mu}{L}\right)^{1_{\{\gamma_s<1\}}} \left(1-\frac{\|g_s\|}{2\|v_s - x_s\|}\right)^{1_{\{\gamma_s=1\}}}$$

where $v_s \in \arg\min_{v\in\mathcal{V}} \langle \nabla f(x_s), v\rangle$ $\forall$ s $\in [T-1]$

Let $\varepsilon_t := f(x_t) - \min_{\mathcal{C}} f$

$$\eta_t = align\left(-\nabla f(x_t), g_t\right) = \frac{\langle -\nabla f(x_t), g_t\rangle}{\|\nabla f(x_t)\|\|g_t\|} \in [0,1] \ \forall\, t \in [0,T]$$

(1) Suppose that $\gamma_t = \eta_t \|\nabla f(x_t)\| / (L\|g_t\|)$

$$\begin{aligned}\varepsilon_{t+1} &\leq \varepsilon_t + \gamma_t\langle \nabla f(x_t), g_t\rangle + \frac{L}{2}\gamma_t^2\|g_t\|^2\\ &= \varepsilon_t - \gamma_t\eta_t\|\nabla f(x_t)\|\|g_t\| + \frac{L}{2}\gamma_t^2\|g_t\|^2\\ &= \varepsilon_t - \eta_t^2\frac{\|\nabla f(x_t)\|^2}{2L}\\ &\leq \varepsilon_t - \eta_t^2\frac{2\mu\varepsilon_t}{2L}\\ &= \left(1-\eta_t^2\frac{\mu}{L}\right)\varepsilon_t\end{aligned}$$

(2) Else $\eta_t \|\nabla f(x_t)\| / (L\|g_t\|) > 1$ and $\gamma_t = 1$

$$\begin{aligned}1 &< \frac{\eta_t\|\nabla f(x_t)\|}{L\|g_t\|}\\ &= \frac{\langle -\nabla f(x_t), g_t\rangle}{L\|g_t\|^2}\\ L\|g_t\|^2 &< \langle -\nabla f(x_t), g_t\rangle\end{aligned}$$

(3) Hence $\varepsilon_{t+1} \leq \varepsilon_t + \gamma_t\langle \nabla f(x_t), g_t\rangle + \frac{L}{2}\gamma_t^2\|g_t\|^2$

$$\begin{aligned}&= \varepsilon_t + \langle \nabla f(x_t), g_t\rangle + \frac{L}{2}\|g_t\|^2\\ &< \varepsilon_t + \frac{\langle \nabla f(x_t), g_t\rangle}{2}\end{aligned}$$

Recall that $g_t = d_{K_t}/\Lambda_t$ and $d_1 = \lambda_0(v_0 - x_t)$ where $v_0 \in \arg\min_{v\in\mathcal{V}} \langle \nabla f(x_t), v\rangle$. If $K_t = 1$ then $d_{K_t} = d_1$, else $align(-\nabla f(x_t), d_{K_t}) > align(-\nabla f(x_t), d_1)$, then $\frac{\langle -\nabla f(x_t), g_t\rangle}{\|g_t\|} \geq \frac{\langle -\nabla f(x_t), v_0 - x_t\rangle}{\|v_0 - x_t\|}$

Let $x^* \in \arg\min_{\mathcal{C}} f$

$$\begin{aligned}\varepsilon_t &= f(x_t) - f(x^*)\\ &\leq \langle \nabla f(x_t), x_t - x^*\rangle\\ &\leq \langle \nabla f(x_t), x_t - v_0\rangle\end{aligned}$$

Thus $\varepsilon_{t+1} < \left(1 - \frac{\|g_t\|}{2\|v_0 - x_t\|}\right)\varepsilon_t$

Together we got

$$f(x_t) - \min_{\mathcal{C}} f \leq \frac{LD^2}{2}\prod_{s=0}^{t-1}\left(1-\eta_s^2\frac{\mu}{L}\right)^{1_{\{\gamma_s<1\}}} \left(1-\frac{\|g_s\|}{2\|v_s - x_s\|}\right)^{1_{\{\gamma_s=1\}}}$$

3.2.2. 推导分析

对 (1):

$0 \to 1$:

$$\begin{aligned}\varepsilon_{t+1} - \varepsilon_t &= f(x_{t+1}) - f(x_t)\\ \text{(L-smooth)} &\leq \varepsilon_t + \langle \nabla f(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2\\ &= \varepsilon_t + \gamma_t\langle \nabla f(x_t), g_t\rangle + \frac{L}{2}\gamma_t^2\|g_t\|^2\end{aligned}$$

$1 \to 2$:

$$\eta_t = \frac{\langle -\nabla f(x_t), g_t\rangle}{\|\nabla f(x_t)\|\|g_t\|}$$

$\langle \nabla f(x_t), g_t\rangle = \eta_t\|\nabla f(x_t)\|\|g_t\|$

$2 \to 3$:

带入 η_t

$3 \to 4$:

$$\frac{\|\nabla f(x)\|^2}{2\mu} \geq 2\mu(f(x_{t+1}) - f(x^*)) = 2\mu\varepsilon_t$$

对 (2):

$0 \to 1$:

由η_t 定义

对 (3):

$0 \to 3$:

对ε_t 推导同 (1) , η_t 取 1

$3 \to 4$:

align 值由 algorithm 保证单调递增

$$\frac{\langle -\nabla f(x_t), \Lambda_t g_t\rangle}{\|\nabla f(x_t)\|\|\Lambda_t g_t\|} \geq \frac{\langle -\nabla f(x_t), v_0 - x_t\rangle}{\|\nabla f(x_t)\|\|v_0 - x_t\|}$$

$$\frac{\langle -\nabla f(x_t), g_t\rangle}{\|g_t\|} \geq \frac{\langle -\nabla f(x_t), v_0 - x_t\rangle}{\|v_0 - x_t\|}$$

$5 \to 7$:

$$\begin{aligned}\text{(convexity of f) } \varepsilon_t &= f(x_t) - f(x^*)\\ \text{(Lemma) } &\leq \langle \nabla f(x_t), x_t - x^*\rangle\\ &\leq \langle \nabla f(x_t), x_t - v_0\rangle\end{aligned}$$

最后一条由两种不同 η_t 的取值所推出的结论合并而来，其中角标表示在范围内取 1，否则取 0

3.3. Worst-case rate

3.3.1. 定理及推导

Theorem5 Let $f : \mathcal{H} \to R$ be L-smooth, convex, and μ-gradient dominated, and set $\gamma_t \leftarrow min\{\eta_t\|\nabla f(x_t)\|/(L\|g_t\|), 1\}$ or $\gamma_t \leftarrow \arg\max_{\gamma\in[0,1]} f(x_t + \gamma g_t)$.Consider Algorithm2 with the minor adjustment $x_{t+1} \leftarrow x_t + \gamma_t'(v_t - x_t)$ in Line (12) when $\gamma_t = 1$, where $v_t \leftarrow v_{k=0}$ is computed in Line(4) and$\gamma_t' \leftarrow min\{\eta_t\|\nabla f(x_t)\|/(L\|g_t\|), 1\}$ or $\gamma_t' \leftarrow \arg\max_{\gamma\in[0,1]} f(x_t + \gamma g_t)$.Then $\forall$ t $\in [0, T]$

$$f(x_t) - \min_{\mathcal{C}} f \leq \frac{4LD^2}{t+2}$$

(1) Let $x^* \in \arg\min_{\mathcal{C}} f$, when t=0

$$\begin{aligned}\varepsilon_0 &= f(x_0) - f(x^*)\\ &\leq f(y) + \langle \nabla f(y), x_0 - y\rangle + \frac{L}{2}\|x_0 - y\|^2 - f(x^*)\\ &\leq f(y) + \langle \nabla f(y), x^* - y\rangle + \frac{LD^2}{2} - f(x^*)\\ &\leq f(x^*) + \frac{LD^2}{2} - f(x^*)\\ &= \frac{LD^2}{2} \leq \frac{4LD^2}{0+2}\end{aligned}$$

Suppose $\varepsilon_t \leq \frac{4LD^2}{t+2},\ \forall\ t \in [0, T]$

If γ_t <1 then we can proceed as in the proof of Theorem 4 and obtain

$$\varepsilon_{t+1} \leq \varepsilon_t - \eta_t^2 \frac{\|\nabla f(x_t)\|^2}{2L}$$

Also there exists $v_t \in \arg\min_{v\in\mathcal{V}} \langle \nabla f(x_t), v\rangle$ such that

$$\begin{aligned}\eta_t &\geq \frac{\langle \nabla f(x_t), x_t - v_t\rangle}{\|\nabla f(x_t)\|\,\|x_t - v_t\|}\\ \varepsilon_t &= f(x_t) - f(x^*)\\ &\leq \langle \nabla f(x_t), x_t - x^*\rangle\\ &\leq \langle \nabla f(x_t), x_t - v_t\rangle\\ \varepsilon_{t+1} &\leq \varepsilon_t - \frac{\langle \nabla f(x_t), x_t - v_t\rangle^2}{\|\nabla f(x_t)\|^2\|x_t - v_t\|^2}\frac{\|\nabla f(x_t)\|^2}{2L}\\ &\leq \varepsilon_t - \frac{\varepsilon_t^2}{D^2}\frac{1}{2L}\\ &= \varepsilon_t\left(1 - \frac{\varepsilon_t}{2LD^2}\right)\end{aligned}$$

If $\varepsilon_t \leq 2LD^2/(t+2)$, then

$$\begin{aligned}\varepsilon_{t+1} &\leq \varepsilon_t\\ &\leq \frac{2LD^2}{t+2}\\ &\leq \frac{4LD^2}{t+3}\end{aligned}$$

(2) Else, $2LD^2/(t+2) < \varepsilon_t \leq 4LD^2/(t+2)$ so

$$\begin{aligned}\varepsilon_{t+1} &\leq \varepsilon_t\left(1 - \frac{\varepsilon_t}{2LD^2}\right)\\ &< \frac{4LD^2}{t+2}\left(1 - \frac{1}{2LD^2}\frac{2LD^2}{t+2}\right)\end{aligned}$$

$$= \frac{4LD^2}{t+2}\left(1-\frac{1}{t+2}\right)$$
$$\leq \frac{4LD^2}{t+2}$$

so $\varepsilon_{t+1} \leq \frac{4LD^2}{t+3}$ holds for t+1 under $\gamma_t < 1$

(3) Now consider the case $\gamma_t = 1$. Then by assumption, $x_{t+1} = x_t + \gamma_t' (v_t - x_t)$

$$\varepsilon_{t+1} \leq \varepsilon_t + \gamma_t' \langle \nabla f(x_t), v_t - x_t \rangle + \frac{L}{2}\gamma_t'^2 \|v_t - x_t\|^2$$

If $\gamma_t' = \langle \nabla f(x_t), x_t - v_t \rangle / \left(L\|x_t - v_t\|^2\right)$ then

$$\varepsilon_{t+1} \leq \varepsilon_t + \gamma_t' \langle \nabla f(x_t), v_t - x_t \rangle + \frac{L}{2}\gamma_t'^2 \|v_t - x_t\|^2$$
$$= \varepsilon_t - \frac{\langle \nabla f(x_t), x_t - v_t \rangle^2}{2L\|x_t - v_t\|^2}$$
$$\leq \varepsilon_t \left(1 - \frac{\varepsilon_t}{2LD^2}\right) \text{ same form proved in (2)}$$

If $\gamma_t' = 1$ then $\langle \nabla f(x_t), x_t - v_t \rangle / (L\|x_t - v_t\|^2) \geq 1$

$$\varepsilon_{t+1} \leq \varepsilon_t + \gamma_t' \langle \nabla f(x_t), v_t - x_t \rangle + \frac{L}{2}\gamma_t'^2 \|v_t - x_t\|^2$$
$$= \varepsilon_t + \langle \nabla f(x_t), v_t - x_t \rangle + \frac{L}{2}\|v_t - x_t\|^2$$
$$\leq \varepsilon_t + \frac{\langle \nabla f(x_t), v_t - x_t \rangle}{2}$$
$$\leq \frac{\varepsilon_t}{2}$$
$$\leq \frac{2LD^2}{t+2}$$
$$\leq \frac{4LD^2}{t+3}$$

So by induction the proof have been completed

$$f(x_t) - \min_{\mathcal{C}} f = \varepsilon_{t+1} - \varepsilon_t \leq \frac{4LD^2}{t+2}$$

3.3.2. 推导分析

对 (1):

$1 \to 2$:

By L-smooth and y $\in \mathcal{C}$

$2 \to 3$:

$$x_0 = \arg\min_{v \in \mathcal{V}} \langle \nabla f(x_t), v \rangle$$
$$\langle \nabla f(y), x_0 \rangle \leq \langle \nabla f(y), x^* \rangle \text{ (Lemma)}$$

$3 \to 4$:

$$f(x^*) \leq f(y) + \langle \nabla f(y), x^* - y \rangle \text{ (convexity of f)}$$

$5 \to 6$:

See proof in Theorem4 (1)

$6 \to 7$:

$$\eta_t = \text{align}(-\nabla f(x_t), g_t)$$
$$= \frac{\langle -\nabla f(x_t), g_t \rangle}{\|\nabla f(x_t)\| \|g_t\|}$$
$$= \frac{\langle -\nabla f(x_t), \Lambda_t g_t \rangle}{\|\nabla f(x_t)\| \|\Lambda_t g_t\|}$$
$$= \text{align}(-\nabla f(x_t), d_{K_t})$$
$$\geq \text{align}(-\nabla f(x_t), v_t - x_t) \ \exists\ v_t \in \mathcal{C} \text{ by algorithm}$$
$$= \frac{\langle -\nabla f(x_t), v_0 - x_t \rangle}{\|\nabla f(x_t)\| \|v_0 - x_t\|}$$

$8 \to 9$:

$$f(x_t) \leq f(x^*) + \langle \nabla f(x^*), x_t - x^* \rangle \text{ (convexity of f)}$$

$9 \to 10$:

$$v_t = \arg\min_{v \in \mathcal{V}} \langle \nabla f(x_t), v \rangle$$
$$\langle \nabla f(x_t), v_t \rangle \leq \langle \nabla f(y), x^* \rangle \text{ (Lemma)}$$

$11 \to 13$:

just put the result proved above together

$13 \to 14$:

$$\varepsilon_{t+1} \leq \varepsilon_t \left(1 - \frac{\varepsilon_t}{2LD^2}\right) \leq \varepsilon_t$$

对 (2):

$3 \to 4$:

$$(t+1)(t+3) < (t+2)(t+2)$$

对 (3):

$0 \to 1$:

$$\varepsilon_{t+1} - \varepsilon_t = f(x_{t+1}) - f(x_t)$$
$$\leq \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2$$
$$\leq \gamma_t' \langle \nabla f(x_t), v_t - x_t \rangle + \frac{L}{2}\gamma_t'^2 \|v_t - x_t\|^2 \text{ (L-smooth)}$$

$3 \to 4$:

$$\varepsilon_t = f(x_t) - f(x^*)$$

$$\leq \langle \nabla f(x^*), x_t - x^* \rangle \text{ (convexity of f)}$$
$$\leq \langle \nabla f(x^*), x_t - v_t \rangle \text{ (Lemma)}$$

$7 \to 8$:

$$0 \geq \langle \nabla f(x^*), v_t - x_t \rangle \text{ (Lemma)}$$

总结：定理在条件中对特殊取值的步长进行了调整，当 $\varepsilon_t = 1$ 时选择了退化为 Frank-Wolfe 的情形 ($v_t = v_0$) 进行收敛性证明。在证明中通过数学归纳法，首先证明 $k = 0$ 时假设成立，再证明在不同步长情况下假设 $k = tc$ 成立，$k = t+1$ 也成立，从而完成证明。

3.4. practical rate

3.4.1. 定理及推导

Theorem6 Let $f : \mathcal{H} \to R$ be L-smooth, convex, and μ-gradient dominated, and set $\gamma_t \leftarrow min\{\eta_t \|\nabla f(x_t)\|/(L\|g_t\|), 1\}$ or $\gamma_t \leftarrow \arg\max_{\gamma \in [0,1]} f(x_t + \gamma g_t)$.Assume that$|\{s \in [0, t-1] \mid \gamma_s < 1, K_s > 1\}| \geq \omega t^p$, $\forall t \in [0, T-1]$, for some $\omega > 0$ and $p \in [0, 1]$.Then $\forall t \in [0, T-1]$

$$f(x_t) - \min_{\mathcal{C}} f \leq \frac{LD^2}{2} \exp\left(-\delta^2 \frac{\mu}{L} \omega t^p\right)$$

denote $N_t := |\{s \in [0, t-1] \mid \gamma_s < 1, K_s > 1\}| \geq \omega t^p$

$$f(x_t) - \min_{\mathcal{C}} f \leq \frac{LD^2}{2} \prod_{s=0}^{t-1} \left(1 - \eta_s^2 \frac{\mu}{L}\right)^{1_{\{\gamma_s < 1\}}} \left(1 - \frac{\|g_s\|}{2\|v_s - x_s\|}\right)^{1_{\{\gamma_s = 1\}}}$$
$$\leq \frac{LD^2}{2} \prod_{\substack{s=0 \\ \gamma_s < 1 \ K_s < 1}}^{t-1} \left(1 - \eta_s^2 \frac{\mu}{L}\right)$$
$$\leq \frac{LD^2}{2} \left(1 - \delta^2 \frac{\mu}{L}\right)^{N_t}$$
$$\leq \frac{LD^2}{2} \exp\left(-\delta^2 \frac{\mu}{L} N_t\right)$$
$$\leq \frac{LD^2}{2} \exp\left(-\delta^2 \frac{\mu}{L} \omega t^p\right)$$

3.4.2. 推导分析

$2 \to 3$:

If $K_t > 1$, then $\eta_s \geq \delta$ by algorithm

$3 \to 4$:

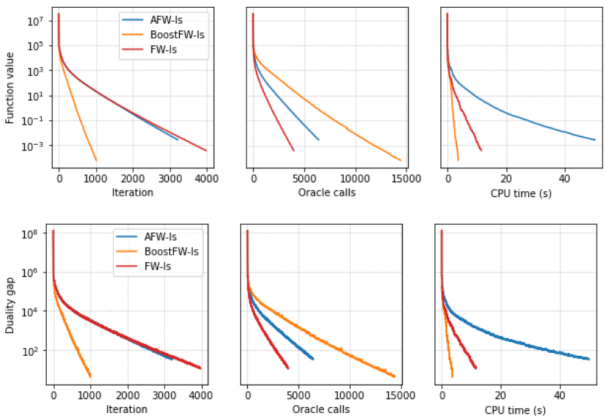
Can be proved if $(1-y)^x \leq (e^{-y})^x$, $\forall\, y \in (0, 1)$

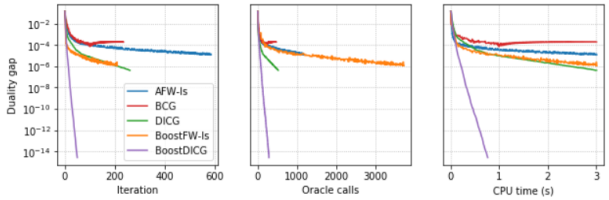
this can be proved if $e^{-y} \geq 1 - y$, $\forall\, y \in (0, 1)$

which is easy to prove

总结：实际收敛速率的证明所增加的条件为在完整过程中满足步长 $\gamma_s < 1$ 且使用 Boosting 拟合负梯度超过一次的迭代点 x_t 有一定的数量 (也即假设 Boosting procedure 被频繁使用)，这也是符合实际实验情况的。

4. **实验复现**





可以看到 Boosting Frank-Wolfe 在大部分的问题上

都要优于同类型改进的 Frank-Wolfe 算法，并且在其他使用了拟合负梯度寻找迭代方向该步骤的类 Frank-Wolfe 算法如 DICG 算法也可对其进行 Boosting 并且表现优于原 DICG 算法。其他的数值实验可在代码附件中找到。

5. 意义及改进

5.1. 算法意义

Boosting Frank-Wolfe 算法对 Frank-Wolfe 算法中寻找迭代方向一步进行了改进，能够更好的拟合负梯度方向，保障了每步迭代的下降量。我们认为本算法的意义在于:

1.Boosting procedure 较大的提升了 Frank-Wolfe 算法的效率，在其他使用了拟合负梯度寻找迭代方向该步骤的类 Frank-Wolfe 算法中也表现良好有，一定应用价值，从数值实验中可以看到

2. 验证了在凸问题中通过保障迭代方向较好拟合负梯度方向从而保证迭代效率是一个比较有效的想法，此想法在非凸问题中保障迭代方向较好拟合-subgradient 说不定也有不错的效果

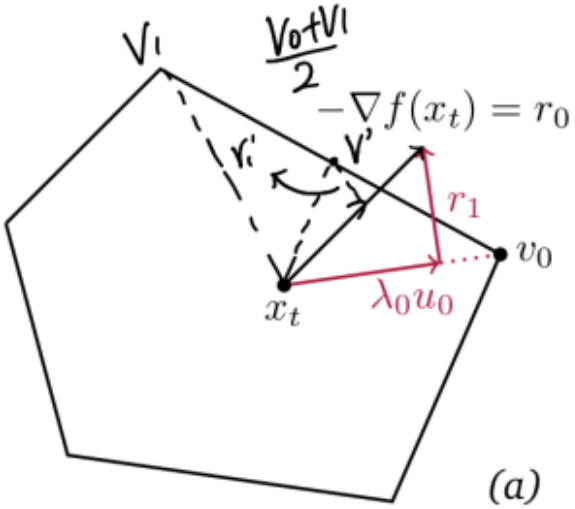
3. 基于 Frank-Wolfe 算法相同的信息 $(\mathcal{V},\nabla f(x_t))$ 进行投影拟合是一个不错的想法，因为这样一种投影能够通过公式直接给出，这种投影方式是否也能适用于其他算法

4. 仅针对 Frank-Wolfe 算法，选择顶点投影的想法还能能够保障迭代点始终在可行域内部，说明类似基于顶点计算迭代方向的想法可能能够有同样良好的性质。

5.2. 算法改进

1. 基于实验观察，同一迭代点 x_t 使用 Boosting procedure 的轮数过多，Boosting procedure 的计算量是比较大的，因为需要计算 $v_k \leftarrow \arg\max_{v\in\mathcal{V}} \langle r_k, v\rangle$，其问题出在无法保证每轮拟合的残差 r 足够小，或是 $align_{t+1} - align_t$ 即角度改进足够多。

目前想到的可能改进方式为：考虑上文例子 可以观察到原本选择的投影方向 $v_0 - x_t$ 所导致的残差 r_1 非常的大，拟合方向与负梯度方向的夹角也很大，其



余顶点也不尽如人意。但是因为可行域是凸集，拥有任意两点连线也属于可行域的性质，因此可以考虑设计一种算法在每轮 Boosting 时根据已有顶点计算新的顶点，如本例中若使用 $\frac{v_0+v_1}{2}$ 作为另时新顶点，那么 r_1' 则会远小于原本的残差。其实更重要的是衡量拟合方向与负梯度的夹角，越小说明拟合越好。

该想法的可行性在于：

1. 基于每轮 Boosting 已经计算了一轮 $v_k \leftarrow \arg\max_{v\in\mathcal{V}} \langle r_k, v\rangle$，所以 $\langle r_k, v\rangle$，$\forall v \in \mathcal{C}$ 是已知的，对于制造新顶点提过了非常多的信息。

2. 若顶点非常多如 1000 次，$v_k \leftarrow \arg\max_{v\in\mathcal{V}} \langle r_k, v\rangle$ 步需要计算一千次内积，因此如果能够通过增加额外几十次内积的计算缩减一次 Boosting，那么也将减少很多计算量，因此此方式值得尝试。

3. 根据 Boosting Frank-Wolfe 证明 $x_{t+1} \in \mathcal{C}$, 可以易证此方法也能保证 $x_{t+1} \in \mathcal{C}$, 因为新顶点是原顶点的线性组合且系数和为 1。

因此该想法的问题就是如何选择 v 满足 $\lambda v + (1-\lambda)v_t$ 甚至更多顶点的线性组合能够逐步减少 $align$(角度)。

算法中可以设置当每轮 $r_k/\lambda_k u_k < \delta$(即迭代方向与负梯度方向的角度) 时才作为拟合方向。由于时间限制，暂未想出一个较好的算法解决该问题，但是我们觉得这是一个值得尝试的方向。