Accelerating Frank-Wolfe via Averaging Step Directions

Zhaoyue Chen
zhaoychen@cs.stonybrook.edu

Yifan Sun
ysun@cs.stonybrook.edu

Abstract

The Frank-Wolfe method is a popular method in sparse constrained optimization, due to its fast per-iteration complexity. However, the tradeoff is that its worst case global convergence is comparatively slow, and importantly, is fundamentally slower than its flow rate–that is to say, the convergence rate is throttled by discretization error. In this work, we consider a modified Frank-Wolfe where the step direction is a simple weighted average of past oracle calls. This method requires very little memory and computational overhead, and provably decays this discretization error term. Numerically, we show that this method improves the convergence rate over several problems, especially after the sparse manifold has been detected. Theoretically, we show the method has an overall global convergence rate of $O(1/k^p)$, where $0 < p < 1$; after manifold identification, this rate speeds to $O(1/k^{3p/2})$. We also observe that the method achieves this accelerated rate from a very early stage, suggesting a promising mode of acceleration for this family of methods.

1 Introduction

The Frank-Wolfe (FW) method (or conditional gradient method) [Dunn and Harshbarger, 1978, Frank et al., 1956] solves the constrained optimization problem

$$\underset{x\in\mathcal{D}}{\text{minimize}} \quad f(x) \tag{1}$$

via the repeated iterations

$$\begin{aligned} \mathbf{s}_k &= \underset{s\in\mathcal{D}}{\operatorname{argmin}} \nabla f(\mathbf{x}_k)^T\mathbf{s}, \\ \mathbf{x}_{k+1} &= \mathbf{x}_k + \gamma_k(\mathbf{s}_k - \mathbf{x}_k). \end{aligned} \tag{FW}$$

Here, we assume that $\mathcal{D}$ is a compact, convex set and $f:\mathbb{R}^n\to\mathbb{R}$ is an L-smooth μ-strongly convex function. The operation in the first line is often referred to as the *linear minimization oracle (LMO)*, and will be expressed succinctly as $\mathbf{s}_k =: \mathrm{LMO}_{\mathcal{D}}(\mathbf{x}_k)$. The FW method has been used in many machine learning applications, such as compressed sensing, recommender systems [Freund et al., 2017], image and video co-localization [Joulin et al., 2014], etc.

A major area of research concerns the slow convergence rate of the FW method. In particular, using a fixed step size sequence $\gamma_k = O(1/k)$, in general the FW method does not converge faster than $O(1/k)$ in objective value. It has been surmised that this is caused by *zig-zagging*, e.g. periodic erratic behavior of the terms $\mathbf{s}_k$, causing $\mathbf{x}_k$ to have bad step directions even arbitrarily close to the solution. For example, this phenomenon was explored in Chen et al. [2021], which showed that in its *continuous* form, the method does not zigzag and also reaches an arbitrarily high convergence rate.

Preprint. Under review.

Moreover, even better multistep discretization methods cannot close this gap. Overall, the conclusion was reached that the largest contributer to the slow convergence rate of FW is the discretization error, which may be reduced as a function of Δ (the discretization step length), but not k (the iteration counter).

1.1 Assumptions and contributions

In this work, we investigate the discretization error term more carefully, and construct an *averaged FW method* as a simple approach to reduce zig-zagging and decay this key discretization term.

Assumptions. Throughout this work, important assumptions are: compact and convex $\mathcal{D}$ (so that the LMO is consistent) and L-smooth f (so that gradients are computable and do not diverge wildly). The extra assumption of convexity is required for the global convergence rate, but is not needed for method implementation. Additionally, μ-strong convexity is only needed for the accelerated local rate, and we believe is a proof artifact; in practice, we observe $\mu = 0$ achieves similar speedups.

Contributions. Overall, for strongly convex objectives, if averaging is done with coefficients $(\frac{c}{c+k})^p$ for constants $c \geq 3p/2 + 1$, $0 < p < 1$, we show

- global convergence of $O(1/k^p)$ for the method and $O(1/t^{1-p})$ for the flow, and
- local convergence of $O(1/k^{3p/2})$ for the method and $O(\log(t)/t^c)$ for the flow.

The differentiation between local and global convergence is characterized by the identification of the sparse manifold ($\bar{k}$, where $\text{LMO}(\mathbf{x}_k) = \text{LMO}(\mathbf{x}^*)$ for all $k \geq \bar{k}$) [Hare and Lewis, 2004, Sun et al., 2019]; this is in similar spirit to works like Liang et al. [2014], which characterize this behavior in proximal gradient methods and other sparsifying methods. Overall, these results suggest improved behavior for sparse optimization applications, which is the primary beneficiary of this family of methods.

1.2 Related works

Accelerated FW. A notable work that highlights the notorious zig-zagging phenomenon is Lacoste-Julien and Jaggi [2015], where an Away-FW method is proposed that cleverly removes offending atoms and improves the search direction. Using this technique, the method is shown to achieve linear convergence under strong convexity of the objective. The tradeoff, however, is that this method requires keeping past atoms, which may incur an undesired memory cost. A work that is particularly complementary to ours is Garber and Hazan [2015], which show an improved $O(1/k^2)$ rate when the *constraint set* is strongly convex–this reduces zigzagging since solutions cannot lie in low-dimensional "flat facets". Our work addresses the exact opposite regime, where we take advantage of "flat facets" in sparsifying sets (1-norm ball, simplex, etc). This allows the notion of *manifold identification* as determining when suddenly the method behavior improves.

Averaged FW. Several previous works have investigated *gradient averaging* [Zhang et al., 2021, Abernethy and Wang, 2017]. While performance seems promising, the rate was not improved past $O(\frac{1}{k})$. Ding et al. [2020] investigates *oracle averaging* by solving small subproblems at each iteration to achieve optimal weights. The work Chen et al. [2021] can also be viewed as an oracle averaging technique, where the weights are chosen inspired by higher order discretization methods. In comparison, in this work the averaging is *intentionally unintelligent*, and the goal is to see how much benefit can be gained simply through this basic approach.

Game theory and fictitious play. A similar version of this method appeared in Hofbauer et al. [2009]. Here, Frank-Wolfe methods are shown to be an instance of a Best Response method over linearized function information at each step. Furthermore, if we imagine players having partial information and instead drawing from a distribution of best responses, this becomes fictitious play, which results in an averaged LMO method. Indeed, our averaged FW method can be viewed as a Best Response Fictitious Play over a uniform distribution of all past LMOs. Therefore, the convergence results in this paper can therefore be applied to the Best Response Fictitious Play as well.

Sparse optimization Other works that investigate *local convergence behavior* include Liang et al. [2014, 2017], Poon et al. [2018]. Here, problems which have these two-stage regimes are described as having *partial smoothness*, which allows for the low-dimensional solution manifold to have significance. Other works that investigate manifold identification include Sun et al. [2019], Iutzeler and Malick [2020], Nutini et al. [2019], Hare and Lewis [2004]. We investigate manifold identification and local convergence of Frank-Wolfe as follows. Since zig-zagging often appears when the solution is on a low-dimensional manifold, we differentiate a local convergence regime of when this manifold is "identified", e.g. all future LMOs are drawn from vertices of this specific manifold. After this point, we show that convergence of both the proposed flow and method can be improved with a decaying discretization error, which in practice may be fast.

2 The problem with vanilla Frank-Wolfe

In order to motivate the new averaged method, let us quickly review the convergence proof for FW method and flow, below

$$\begin{aligned} s(t) &= \mathrm{LMO}_{\mathcal{D}}(-\nabla f(x(t)) \\ \dot{x}(t) &= \gamma(t)(s(t)-x(t)) \end{aligned} \quad \text{(FWFlow)}, \qquad \begin{aligned} \mathbf{s}_k &= \mathrm{LMO}_{\mathcal{D}}(-\nabla f(\mathbf{x}_k) \\ \mathbf{x}_{k+1} &= \mathbf{x}_k + \gamma_k(\mathbf{s}_k - \mathbf{x}_k) \end{aligned} \quad \text{(FW)}.$$

Specifically, the FWFlow [Jacimovic and Geary, 1999] is the continuous-time trajectory for which the FW [Frank et al., 1956, Dunn and Harshbarger, 1978] is its Euler's explicit discretization. Note that flows are not an implementable method, but rather an analysis tool that bypasses discretization artifacts.

Consider now the trajectory of the objective flow loss $h(t) := f(x(t)) - \min_{x\in\mathcal{D}} f(x)$. Using the properties of the LMO and convexity of f, it can be shown that

$$\dot{h}(t) = \nabla f(x)^T \dot{x}(t) = \gamma(t)\nabla f(x)^T(s-x) \leq -\gamma(t)h(t) \tag{2}$$

and taking $\gamma(t) = \frac{c}{c+t}$, we arrive at

$$h(t) \leq \frac{h(0)}{(c+t)^c} = O(1/t^c).$$

In contrast, using L-smoothness, the FW method satisfies the recursion

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \quad \leq \quad \gamma_k \underbrace{\nabla f(\mathbf{x}_k)^T(\mathbf{x}_k - \mathbf{s}_k)}_{\leq -\mathbf{gap}(x)} + \frac{L\gamma_k^2}{2} \underbrace{\|\mathbf{x}_k - \mathbf{s}_k\|_2^2}_{=(2D)^2}$$

where $D = \max_{x\in\mathcal{D}} \|x\|_2$ and the duality gap $\mathbf{gap}(\mathbf{x}) \geq f(\mathbf{x}) - f(\mathbf{x}^*)$ for all convex f. Defining $\mathbf{h}_k = \mathbf{h}(\mathbf{x}_k)$

$$\mathbf{h}_{k+1} - \mathbf{h}_k \leq -\gamma_k \mathbf{h}_k + 2LD^2\gamma_k^2 \tag{3}$$

which recursively gives $\mathbf{h}_k = O(\frac{c}{c+k})$. Note the key difference in (3) and its analogous terms in (2) is the extra $2LD^2\gamma_k^2 = O(1/k^2)$ term, which throttles the recursion from doing better than $\mathbf{h}_k = O(1/k)$; in particular, $\mathbf{h}_{k+1} - \mathbf{h}_k \geq \Omega(1/k^2)$. One may ask if it is possible to bypass this problem by simply picking $\gamma_k = O(1/k^p)$ more aggressively, e.g. $p > 1$; however, then such a sequence becomes summable, and convergence of $\mathbf{x}_k \to \mathbf{x}^*$ is not assured. Therefore, we must have a sequence γ_k converging *at least* as slowly as $O(1/k)$.

Thus, the primary culprit in this convergence rate tragedy is the bound $\|\mathbf{s}_k - \mathbf{x}_k\|_2 = O(D)$ (nondecaying). This bound is also not in general loose; if $\mathbf{x}^*$ is in the interior of $\mathcal{D}$, then even at optimum, $\mathbf{s}_k$ may be any vertex in $\mathcal{D}$ and may bounce around the boundary of the set. As an example, consider the 1-D optimization $\min_{-1\leq x\leq 1} x^2$. Then, assuming $\mathbf{x}_0 \neq 0$, $\mathbf{s}_k = \mathbf{sign}(-\mathbf{x}_k)$ for all k, and

$$\|\mathbf{x}_k - \mathbf{s}_k\|_2 = \|\mathbf{x}_k + \mathbf{sign}(\mathbf{x}_k)\|_2 \geq 1, \quad \forall k.$$

That is to say, this error term *does not decay in general, even when* $\mathbf{x}_k \approx \mathbf{x}^*$.

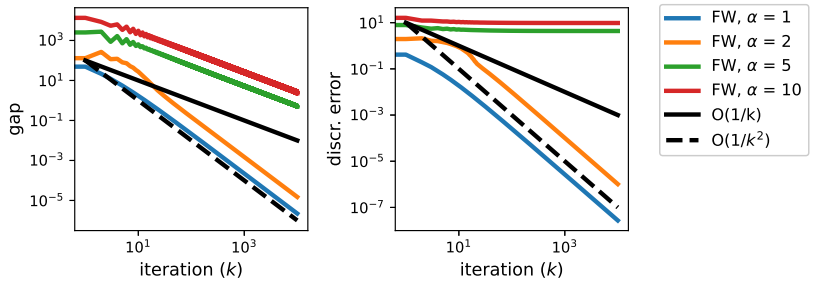


Figure 2: An example of discretization error decay and method improvement when $\mathbf{x}^*$ is on the boundary of a strongy convex set (in this case, the 2-norm ball). Unconstrained, $\|\mathbf{x}^*\|_2 \approx 2.44$.

Specifically, consider $\mathcal{S}(\mathbf{x}^*)$ the set of all possible LMOs at $\nabla f(\mathbf{x}^*)$. [1] There are three possible scenarios.

- **Case I** (C): $\mathcal{S}(\mathbf{x}^*)$ contains only 1 element. This occurs when $\mathcal{D}$ is a *strongly convex set*, like the 2-norm ball. It can also be a feature in parts of the boundary of other sets, like the 1,2-norm ball, group norm ball, etc.
- **Case II** (F): $\mathcal{S}(\mathbf{x}^*)$ contains multiple elements. This occurs when $\mathbf{x}^*$ is on the boundary of $\mathcal{D}$, which itself is an *atomic* set, e.g. the simplex, the ℓ_1 ball, and linear transformations of these. This also includes the nuclear norm ball and the group norm ball.
- **Case III** (G): $\mathbf{x}^*$ is in the interior of any set, and $\mathcal{S}(\mathbf{x}^*)$ contains multiple elements, whether $\mathcal{D}$ is strongly convex or not.

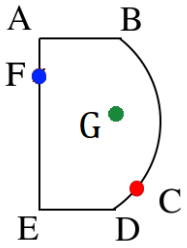


Figure 1: Example of a set $\mathcal{D}$. Here, A, B, D, and E are its isolated extremal vertices; and F and C are two candidate solution points.

These three cases are illustrated in Figure 1; at C, the LMO is necessarily unique, and spits out $\mathbf{s}_k = \mathbf{x}_k$; at F, the LMO may choose vertices A or E, and will oscillate between them forever. Similarly, at G, the LMO will bounce between all points of which G is its convex combination, forever. We quickly discuss Case I, addressed by past works [Garber and Hazan, 2015], where it is shown FW has a $O(1/k^2)$ (accelerated) convergence rate when $\mathbf{x}^*$ is on the boundary of a strongly convex set. The transition from case I to case III is shown in Figure 2, where if α is small (and $\mathbf{x}^*$ is on the boundary of the 2-norm ball) convergence is accelerated (and discretization error decays); on the other hand, when α is large (and $\mathbf{x}^*$ is in the interior), we see no such acceleration. Therefore, in cases II and III, an entirely new method must be defined to reduce this error term.

3 An averaged Frank-Wolfe method

We now propose an LMO-averaged Frank-Wolfe (AvgFW) method, by replacing $\mathbf{s}_k$ with an averaged version $\bar{\mathbf{s}}_k$. The proposed flow and method are as follows:

$$\begin{aligned} s(t) &= \mathrm{LMO}_{\mathcal{D}}(x(t)) \\ \dot{\bar{s}}(t) &= \beta(t)(s(t) - \bar{s}(t)) \quad (\textsc{AvgFWFlow}), \\ \dot{x}(t) &= \gamma(t)(\bar{s}(t) - x(t)) \end{aligned} \qquad\qquad \begin{aligned} \mathbf{s}_k &= \mathrm{LMO}_{\mathcal{D}}(\mathbf{x}_k) \\ \bar{\mathbf{s}}_k &= \bar{\mathbf{s}}_{k-1} + \beta_k(\mathbf{s}_k - \bar{\mathbf{s}}_{k-1}) \quad (\textsc{AvgFW}) \\ \mathbf{x}_{k+1} &= \mathbf{x}_k + \gamma_k(\bar{\mathbf{s}}_k - \mathbf{x}_k) \end{aligned}$$

[1] This set $\mathcal{S}(\mathbf{x}^*)$ is actually the subdifferential $\partial\sigma_{\mathcal{D}}(-\nabla f(\mathbf{x}^*))$, where $\sigma_{\mathcal{D}}(z)$ is the support function of $\mathcal{D}$ at z. Note that this is a well-defined quantity; although we do not require f to be strongly convex (and thus $\mathbf{x}^*$ may not be unique), we do require it to be L-smooth (and thus $\nabla f(\mathbf{x}^*)$ *is* unique).

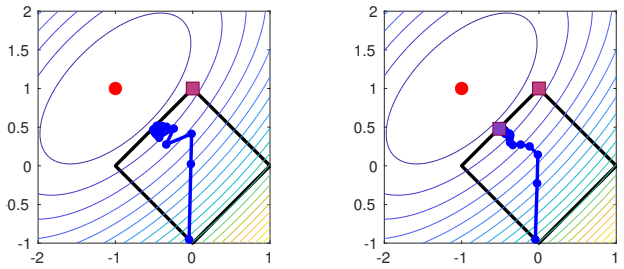


Figure 3: *Top*: FW, with last atom $\mathbf{s}_k$ (light purple). *Bottom*: AVGFW, with last atom $\mathbf{s}_k$ (light purple), and averaged atom $\bar{\mathbf{s}}_k$ (dark purple).

where $\beta_k = (\frac{c}{c+k})^p$ for $c \geq 0$ and $0 < p \leq 1$. [2] Here, the smoothing in $\bar{s}_k$ and $\bar{\mathbf{s}}_k$ has two roles. First, averaging reduces zigzagging; at every k, $\bar{\mathbf{s}}_k$ is a convex combination of past $\mathbf{s}_k$, and has a smoothing effect that qualitatively also reduces zig-zagging; this is of importance should the user wish to use line search or momentum-based acceleration. Second, this forces the discretization term $\|\mathbf{x}_k - \bar{\mathbf{s}}_k\|_2$ to *decay*, allowing for faster numerical performance.

Figure 3 shows a simple 2D example of our AVGFWmethod (right) compared to the usual FWmethod (left). It is clear that the averaged discretization term $\|\bar{\mathbf{s}}_k - \mathbf{x}_k\|_2$ decays, whereas the vanilla discretization term $\|\mathbf{s}_k - \mathbf{x}_k\|_2$ never converges.

4 Convergence analysis

4.1 Global convergence

We start by showing the method converges, with no assumptions on sparsity or manifold identification. While in practice we see faster convergence of AVGFWcompared to FW, averaging produces challenges in the convergence proof.

Specifically, we require that the weights $\beta_k = (\frac{c}{c+k})^p$ be a little skewed toward more recent values; in practice, $p = 1$ works optimally, but in analysis we require $0 < p < 1$. The convergence proof then expands a gap-like residual term as follows:

$$\bar{\mathbf{g}}_k := \nabla f(\mathbf{x}_k)^T(\bar{\mathbf{s}}_k - \mathbf{x}_k) \leq -\beta_k \mathbf{gap}(\mathbf{x}_k) + (1-\beta_k)(1-\gamma_{k-1})\bar{\mathbf{g}}_{k-1} + O(\gamma_k)$$

and from there, form a recursion on the objective error. Unfortunately, that without relieving the $O(\gamma_k^2)$ constant term, this rate cannot possibly be better than that of the vanilla FW. This does contradict our numerical results, however, which show a global accelerated rate as well, suggesting that this rate could be tightened with the right assumptions. However, at this point our goal is to be fully general, so that the rates can then be used to bound manifold identification and discretization error decay.

Theorem 4.1 (Global rates)**.** *Take* $0 \leq p \leq 1$.

- *For* $\beta(t) = \left(\frac{c}{c+t}\right)^p$, *the flow* AVGFWFLOW *satisfies* $f(x(t)) \leq O\left(\frac{1}{t^{1-p}}\right)$.
- *For* $\beta_k = \left(\frac{c}{c+k}\right)^p$, *the method* AVGFW *satisfies* $f(\mathbf{x}_k) \leq O\left(\frac{1}{k^p}\right)$.

The proofs are in Appendix C. Note that the rates are not exactly reciprocal. In terms of analysis, the method is that allows the term $\beta_k \mathbf{gap}(\mathbf{x}_k)$ to take the weight of an entire step, whereas in the flow, the corresponding term is infinitesimally small. This is a curious disadvantage of the flow analysis; since most of our progress is accumulated in the last step, the method exploits this more readily than the flow, where the step is infinitesimally small.

[2] Recall that $\gamma_k = c/(c+k)$. While it is possible to use $\beta(t) = b/(b+t)$ with $b \neq c$ in practice, our proofs considerably simplify when the constants are the same.

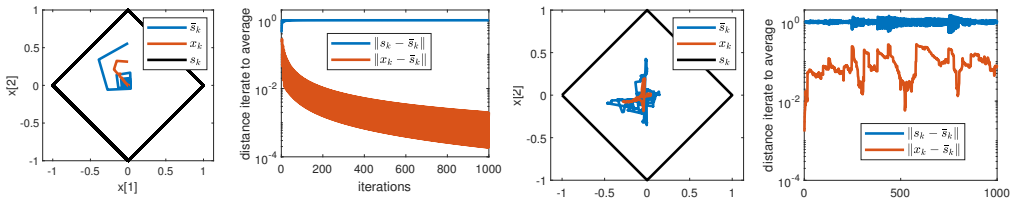


Figure 4: **Two examples over ℓ_1 norm ball.** *Left:* When oscillation is a repeating sequence, the distance $\|\mathbf{x}_k - \bar{\mathbf{s}}_k\|_2$ decays at a $O(1/k)$ rate. *Right:* When oscillation is random, this term does not decay in general.

4.2 Averaging decays discretization term

Now that we have established the convergence rate of the method, we can now use it to bound the decay of the discretization term. There are two possible pitfalls here. First, note that, though $x(t)$ can be viewed as an average of $s(t)$, the fact that $\|\dot{x}\|_2 \to 0$ is not sufficient to suggest that $\bar{s}(t)$ ever stops moving (as is evidenced in the vanilla FW method, where $x(t)$ averages $s(t)$ which may oscillate forever). Conversely, the fact that $\bar{s}(t)$ is an average in itself is also not sufficient; the averaging rate of $O(1/k)$ is not summable, so though $\|\dot{\bar{s}}\|_2 \to 0$, $\bar{s}$ may still never converge. (e.g., $\bar{s}(t) = \log(t)$.)

In fact, we need both properties working in tandem. Intuitively, the fact that $x(t)$ indeed *does* stop moving, halting at some x^*, means that if $\bar{s}(t)$ indeed keeps moving forever, it cannot do so in a divergent path, but must keep orbiting x^*. Combined with the averaging will force $s(t) \to x^*$. [3] We prove this precisely in Appendix B; here, we provide a "proof by picture" (Figure 4) when $\mathbf{s}_k$ follows some prespecified trajectory (unrelated to any optimization problem). In one case, $\mathbf{x}_k$ becomes stationary, and $\bar{\mathbf{s}}_k \to \mathbf{x}_k$; in the other case, $\mathbf{x}_k$ keeps moving forever, suggesting the averages never have the chance to catch up.

4.3 Manifold identification

Now that we know that the method converges and the discretization term collapses, we have an accelerated local convergence rate. Specifically, we consider Case II: $\mathbf{x}^*$ is on the boundary of $\mathcal{D}$, on a low-dimensional facet, and is a combination of multiple (but not all) extremal vertices of $\mathcal{D}$. Specifically, in this case, $\text{LMO}_{\mathcal{D}}(\mathbf{x}^*)$ may return multiple values. This is the usual case in sparse optimization. In this case, we may distinguish between two phases of the trajectory: first identifying the sparse manifold (global convergence), and then iterating after the sparse manifold has been identified (local convergence).

A manifold identification *rate* is expressed as $\bar{k}$, where for any $k > \bar{k}$, any $\text{LMO}_{\mathcal{D}}(\mathbf{x}_k)$ is also an $\text{LMO}_{\mathcal{D}}(\mathbf{x}^*)$. As an example, when $\mathcal{D}$ is the one-norm ball, then at optimality, $\mathbf{x}^*_i > 0$ only if $|\nabla f(\mathbf{x}^*)_i| = \|\nabla f(\mathbf{x}^*)\|_\infty$. We define a *degeneracy parameter* δ as

$$\delta = \min_{j:\mathbf{x}^*_j = 0} \|\nabla f(\mathbf{x}^*)\|_\infty - |\nabla f(\mathbf{x}^*)_j|.$$

If $f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \delta/(Ln)$, then

$$\|\nabla f(x) - \nabla f(x^*)\|_\infty \leq n\|\nabla f(x) - \nabla f(x^*)\|_2 \leq \frac{n}{L}(f(\mathbf{x}_k) - f(\mathbf{x}^*)) \leq \delta/2.$$

Then, any LMO of x must necessarily be a maximizer of $\mathbf{x}^*$. Therefore, a manifold identification upper bound is $\bar{k}$ where for all $k > \bar{k}$, $f(\mathbf{x}_k) - f(\mathbf{x}^*) < \delta/(Ln)$; that is, the rate is the same as the global convergence rate, scaled by problem-dependent parameters.

This simple idea can be extended to any polyhedral set $\mathcal{D}$, where the degeneracy parameter δ will depend on the L-smoothness of f with respect to the gauge function of $\mathcal{D}$ [Freund, 1987, Friedlander et al., 2014]. Concretely, we define $\mathcal{S}(x)$ (the support of x w.r.t. $\mathcal{D}$) as the set of vertices in $\mathcal{D}$ that may be retured by the LMO; that is,

$$\mathcal{S}(x) := \{s : s^T \nabla f(x) = \min_{s' \in \mathcal{D}} s'^T \nabla f(x)\}$$

[3] Intuitively, this is similar to the concept that the alternating sequence $(-1)^k/k$ is summable but $1/k$ is not.

and we say that the manifold has been identified at $\bar{k}$ if for all $k > \bar{k}$, $\text{LMO}(\mathbf{s}_k) \in \mathcal{S}(\mathbf{x}^*)$. A similar δ can be computed in this case. The point is, in all these cases, the manifold identification rate is simply a $1/\delta$ scaling of the method convergence rate. (See also Sun et al. [2019].) [4]

4.4 Accelerated local convergence

To see how averaging affects the convergence rate, first let us consider the flow of AvgFW:

$$\begin{aligned}\frac{\partial}{\partial t} f(x(t)) &= \gamma(t) \nabla f(x(t))^T (\bar{s}(t) - x(t)) \\ &= \gamma(t) \underbrace{\nabla f(x(t))^T (\bar{s}(t) - \hat{s}(t))}_{A} + \gamma(t) \underbrace{\nabla f(x(t))^T (\hat{s}(t) - x(t))}_{B}\end{aligned}$$

where we carefully pick $\hat{s}(t)$ as follows:

$$\dot{\hat{s}}(t) = \beta(t)(\tilde{s}(t) - \hat{s}(t)), \quad \tilde{s}(t) = \operatorname*{argmin}_{s \in \mathcal{S}(x^*)} \|s(t) - s\|_2.$$

In other words, once $\mathcal{S}(x(t)) = \mathcal{S}(x^*)$ (manifold identified), then $\tilde{s}(t) = s(t)$ for all subsequent t. Using this last choice of $\hat{s}(t)$, we are able to ensure A decaying via averaging, and and $B = -\mathbf{gap}(x(t))$ after manifold identification.

Theorem 4.2 (Local rate)**.** *After manifold identification, the flow AvgFWFlow satisfies*

$$f(x(t)) - f(x^*) \leq \frac{\log(t)}{t^c}.$$

The proof is in appendix D.

Now let us extrapolate to the method. From L-smoothness of f, we have the difference inequality

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq \gamma_k \underbrace{\nabla f(\mathbf{x}_k)^T (\bar{\mathbf{s}}_k - \hat{\mathbf{s}}_k)}_{A} + \gamma_k \underbrace{\nabla f(\mathbf{x}_k)^T (\hat{\mathbf{s}}_k - \mathbf{x}_k)}_{B} + \frac{\gamma_k^2 L}{2} \underbrace{\|\bar{\mathbf{s}}_k - \mathbf{x}_k\|_2^2}_{C}.$$

The same tricks work on terms A and B; however, as in the vanilla case, the rate is throttled by $f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \geq O(C)$. We now combine in our averaging result, which, mixed with the global convergence rate (which has no assumptions in discretization decay) shows that $\|\mathbf{s}_k - \mathbf{x}_k\|_2^2 = O(1/k^{3p/2-1})$. This gives our final convergence result.

Theorem 4.3 (Local rate)**.** *Assume that f is μ-strongly convex, and pick $c \geq 3p/2+1$. After manifold identification, the method AvgFW satisfies*

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) = O(1/k^{3p/2}).$$

The proof is in appendix D. Although the proof requires $p < 1$, in practice, we often use $p = 1$ and observe about a $O(1/k^{3/2})$ rate, regardless of strong convexity.

5 Numerical Experiments

5.1 Simulated Compressed Sensing

We minimize a quadratic function with a ℓ_1- norm ball constraint. Given $\mathbf{x}_0 \in \mathbb{R}^m$, a sparse ground truth vector with 10% nonzeros, and given $A \in \mathbb{R}^{n \times m}$ with entries i.i.d. Gaussian, we generate $\mathbf{y} = A\mathbf{x}_0 + \mathbf{z}$ where $\mathbf{z}_i \sim \mathcal{N}(0, 0.05)$. Then we solve

$$\min_{x \in \mathbb{R}^n} \quad \tfrac{1}{2}\|Ax - y\|_2^2 \qquad \text{subject to} \quad \|x\|_1 \leq \alpha. \tag{4}$$

[4] Note that if $\mathcal{D}$ is a strongly convex set, then if $\mathbf{x}^*$ is on the boundary (Case I), then $\mathcal{S}(\mathbf{x}^*)$ is a singleton and moreover $\delta = 0$ and manifold identification cannot happen in finite time. For this reason, this is not a case that benefits from this analysis; we really are only considering polyhedral $\mathcal{D}$. Note that in Case III, manifold identification is simply identified from the very start.

Figure 5 evaluates the performance of FW and AvgFW, for a problem with with $m = 100$, $n = 500$, and varying values of α. Note that this problem is *not* strongly convex. In all cases, we see an improvement in the duality gap (which upper bounds the objective suboptimality $f(\mathbf{x}) - f(\mathbf{x}^*)$) improve its convergence rate from $O(1/k)$ to approaching $O(1/k^{3/2})$, and in all cases the discretization rate of AvgFW is comparable to that of the gap. The support size itself varies based on α, and when α is neither very big nor very small, seems to decay slowly; however, the final sparsity set and the initial working set (number of nonzeros touched from that point on) are very similar, suggesting that though the optimal manifold was not found exactly, it was well-approximated.

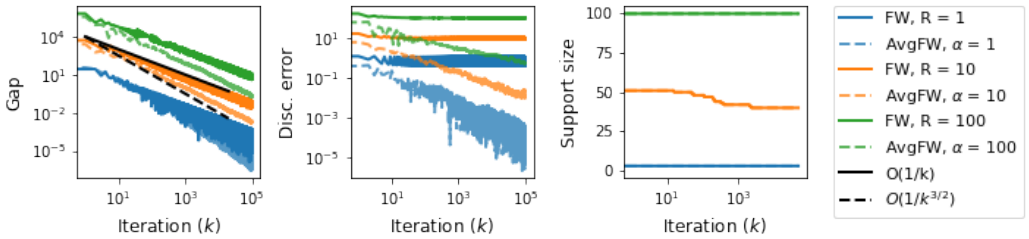


Figure 5: **Compressed sensing.** *Left:* Gap. *Center:* Discretization error, $\|\mathbf{s}_k - \mathbf{x}_k\|_2$ for FW and $\|\bar{\mathbf{s}}_k - \mathbf{x}_k\|_2$ for AvgFW. *Right:* Support size, which is the number of unique indices from current iteration till the end.

5.2 Sparse Logistic Regression

We now evaluate our proposed AvgFW method on two binary classification tasks. We minimize

$$\min_{x \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^{m} \log(1 + \exp(-y_i z_i^T x)) \quad \text{s.t. } \|x\|_1 \leq \alpha \tag{5}$$

for two real world datasets [Guyon et al., 2004].

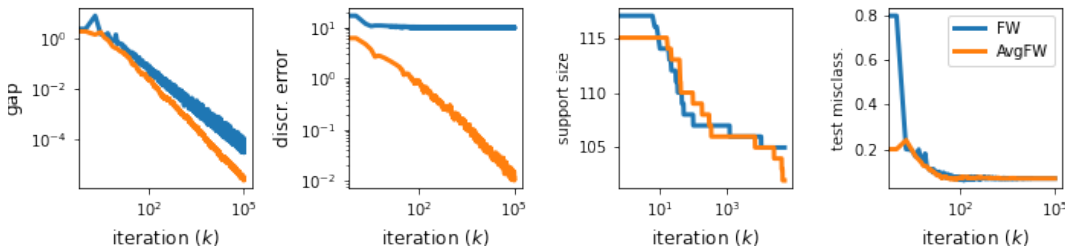
- **Low sparsity problem (Fig 6a).** The Gisette task is to recognize handwritten digits; to use binary classification, we only diffentiate between 4 and 9. The dataset is fully dense, with $n = 5000$ features, so the entry in Gisette dataset is dense. We use a 60/30 train/validation split over $m = 2000$ data samples, and we set $\alpha = 10$.
- **High sparsity problem (Fig 6b).** We consider the task of predicting which chemical compounds bind to Thrombin from the Dorothea drug discovery dataset. The dataset is sparse, with about 0.91% nonzeros, and the labels are unbalanced (9.75% +1, 90.25% -1). We use the given $m = 800$ training samples and $m = 350$ validation samples, with $n = 100000$ features, and $\alpha = 10$.

In both cases, we optimize α over the validation set, sweeping a coarse logarithmic grid of 10 points from 1 to 100. Here, we see similar speedups in the duality gap, discretization error, and support size.

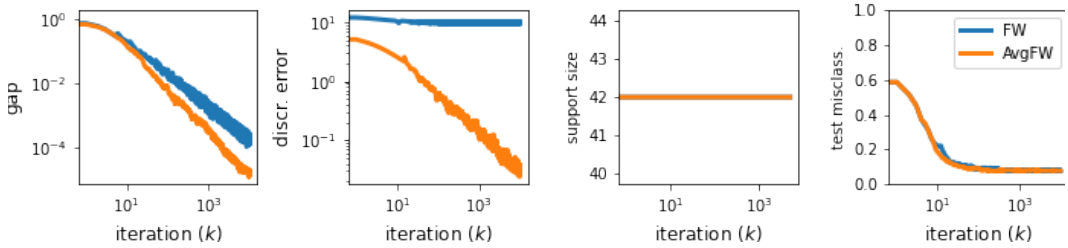
6 Discussion

The main goal of this study is to see if the discretization error, which seems to be the primary cause of slow FW convergence, can be attacked directly using an averaging method, and thereby speed up the convergence rate. This was accomplished; without using away steps we are able to improve the $O(1/k)$ rate to up to $O(1/k^{3/2})$ with negligible computation and memory overhead; moreover, the offending term in the discretization error, which is constant in FW when $\mathcal{D}$ is not strongly convex, is shown here to always decay. A global and local convergence rate is given.

Our numerical results show that, though our theoretical improvements are only local, the effect of this acceleration appears effective globally; moreover, manifold identification (or at least reduction to a small working set of nonzeros) appears almost immediately in many cases. In all cases, we do note that though averaging improves duality gap, it does not seem to appreciably improve



(a) **Gisette.**



(b) **Dorothea.**

Figure 6: Performance of AvgFW on binary classification over real world datasets.

manifold identification or test error performance, suggesting that a fast-and-loose implementation (no averaging) works well for all practical reasons. Still, we motivate that the improvement in the convergence rate is valuable in providing reliability guarantees in downstream applications.

References

Jacob D Abernethy and Jun-Kun Wang. On frank-wolfe and equilibrium computation. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Zhaoyue Chen, Mok Siang Lee, and Yifan Sun. Continuous time frank-wolfe does not zig-zag, but multistep methods do not accelerate. 2021.

Lijun Ding, Jicong Fan, and Madeleine Udell. $k$fw: A frank-wolfe style algorithm with stronger subproblem oracles. *arXiv preprint arXiv:2006.16142*, 2020.

Joseph C Dunn and S Harshbarger. Conditional gradient algorithms with open loop step size rules. *Journal of Mathematical Analysis and Applications*, 62(2):432–444, 1978.

Marguerite Frank, Philip Wolfe, et al. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

R. Freund, Paul Grigas, and R. Mazumder. An extended frank-wolfe method with "in-face" directions, and its application to low-rank matrix completion. *SIAM J. Optim.*, 27:319–346, 2017.

Robert M Freund. Dual gauge programs, with applications to quadratic programming and the minimum-norm problem. *Mathematical Programming*, 38(1):47–67, 1987.

Michael P Friedlander, Ives Macedo, and Ting Kei Pong. Gauge optimization and duality. *SIAM Journal on Optimization*, 24(4):1999–2022, 2014.

Dan Garber and Elad Hazan. Faster rates for the frank-wolfe method over strongly-convex sets. In *International Conference on Machine Learning*, pages 541–549. PMLR, 2015.

Isabelle Guyon, Steve R Gunn, Asa Ben-Hur, and Gideon Dror. Result analysis of the nips 2003 feature selection challenge. In *NIPS*, volume 4, pages 545–552, 2004.

Warren L Hare and Adrian S Lewis. Identifying active constraints via partial smoothness and prox-regularity. *Journal of Convex Analysis*, 11(2):251–266, 2004.

Josef Hofbauer, Sylvain Sorin, and Yannick Viossat. Time average replicator and best-reply dynamics. *Mathematics of Operations Research*, 34(2):263–269, 2009.

Franck Iutzeler and Jérôme Malick. Nonsmoothness in machine learning: specific structure, proximal identification, and applications. *Set-Valued and Variational Analysis*, 28(4):661–678, 2020.

Milojica Jacimovic and Andjelija Geary. A continuous conditional gradient method. *Yugoslav journal of operations research*, 9(2):169–182, 1999.

Armand Joulin, Kevin D. Tang, and Li Fei-Fei. Efficient image and video co-localization with frank-wolfe algorithm. In *ECCV*, 2014.

Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of frank-wolfe optimization variants. In *NIPS*, 2015.

Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Local linear convergence of forward–backward under partial smoothness. *Advances in neural information processing systems*, 27, 2014.

Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Local convergence properties of douglas–rachford and alternating direction method of multipliers. *Journal of Optimization Theory and Applications*, 172(3):874–913, 2017.

Julie Nutini, Mark Schmidt, and Warren Hare. "active-set complexity" of proximal gradient: How long does it take to find the sparsity pattern? *Optimization Letters*, 13(4):645–655, 2019.

Clarice Poon, Jingwei Liang, and Carola Schoenlieb. Local convergence properties of saga/prox-svrg and acceleration. In *International Conference on Machine Learning*, pages 4124–4132. PMLR, 2018.

Yifan Sun, Halyun Jeong, Julie Nutini, and Mark Schmidt. Are we there yet? manifold identification of gradient-related proximal methods. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1110–1119. PMLR, 2019.

Yilang Zhang, Bingcong Li, and Georgios B. Giannakis. Accelerating frank-wolfe with weighted average gradients. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5529–5533, 2021. doi: 10.1109/ICASSP39728.2021.9414485.

A Accumulation terms

Lemma A.1. *For an averaging term $\bar{s}(t)$ satisfying*

$$\dot{\bar{s}}(t) = \beta(t)(s(t) - \bar{s}(t)), \qquad \bar{s}(0) = s(0) = 0$$

where $\beta(t) = \frac{c^p}{(c+t)^p}$, then

$$\bar{s}(t) = \begin{cases} e^{-\alpha(t)} \int_0^t \frac{c^p e^{\alpha(\tau)}}{(c+\tau)^p} s(\tau) d\tau, & p \neq 1 \\ \frac{c}{(c+t)^c} \int_0^t (c+\tau)^{c-1} s(\tau) d\tau & p = 1 \end{cases}$$

where $\alpha(t) = \frac{c^p (c+t)^{1-p}}{1-p}$. If $s(t) = 1$ for all t, then we have an accumulation term

$$\bar{s}(t) = \begin{cases} 1 - \frac{e^{\alpha(0)}}{e^{\alpha(t)}}, & p \neq 1 \\ 1 - (\frac{c}{c+t})^c, & p = 1 \end{cases}$$

Proof. This can be done through simple verification.

- If $p \neq 1$,

  $$\alpha'(t) = \frac{c^p}{(c+t)^p} = \beta(t),$$

  and via chain rule,

  $$\bar{s}'(t) = \underbrace{e^{-\alpha(t)} \frac{c^p \exp(\alpha(t))}{(c+t)^p}}_{=\beta(t)} s(t) - \alpha'(t) \underbrace{\exp(-\alpha(t)) \int_0^t \frac{c^p \exp(\alpha(\tau))}{(c+\tau)^p} s(\tau) d\tau}_{\bar{s}(t)}.$$

  The accumulation term can be verified if

  $$e^{-\alpha(t)} \int_0^t \frac{c^p e^{\alpha(\tau)}}{(c+\tau)^p} d\tau = 1 - \frac{e^{\alpha(0)}}{e^{\alpha(t)}}$$

  which is true since

  $$e^{-\alpha(t)} \int_0^t \frac{c^p e^{\alpha(\tau)}}{(c+\tau)^p} d\tau = e^{-\alpha(t)} \int_0^t (\frac{d}{d\tau} e^{\alpha(\tau)}) d\tau.$$

- If $p = 1$

  $$\bar{s}'(t) = \frac{c}{(c+t)} s(t) - \frac{c^2}{(c+t)^{c+1}} \int_0^t (c+\tau)^{c-1} s(\tau) d\tau = \frac{c}{(c+t)} (s(t) - \bar{s}(t)).$$

  For the accumulation term,

  $$\frac{c}{(c+t)^c} \int_0^t (c+\tau)^{c-1} d\tau = \frac{c}{(c+t)^c} \int_0^t (\frac{\partial}{\partial \tau} \frac{(c+\tau)^c}{c}) d\tau = 1 - (\frac{c}{c+t})^c.$$

□

For convenience, we define

$$\beta_{t,\tau} := \begin{cases} \frac{c^p e^{\alpha(\tau) - \alpha(t)}}{(c+\tau)^p}, & p \neq 1 \\ \frac{c(c+\tau)^{c-1}}{(c+t)^b}, & p = 1, \end{cases} \qquad \bar{\beta}_t := \begin{cases} 1 - \frac{\exp(\alpha(0))}{\exp(\alpha(t))}, & p \neq 1 \\ 1 - (\frac{c}{c+t})^b & p = 1 \end{cases}$$

Lemma A.2. *For the averaging sequence $\bar{\mathbf{s}}_k$ defined recursively as*

$$\bar{\mathbf{s}}_{k+1} = \bar{\mathbf{s}}_k + \beta_k(\mathbf{s}_k - \bar{\mathbf{s}}_k), \qquad \bar{\mathbf{s}}_0 = 0.$$

Then

$$\bar{\mathbf{s}}_k = \sum_{i=1}^{k} \beta_{k,i}\mathbf{s}_i, \qquad \beta_{k,i} = \frac{c^p}{(c+i)^p}\prod_{j=0}^{k-i-1}\left(1 - \frac{c^p}{(c+k-j)^p}\right) \overset{p=1}{=} \frac{c}{c+i}\prod_{j=0}^{c}\frac{i+j+1}{c+k-j}$$

and moreover, $\sum_{i=1}^{k}\beta_{k,i} = 1$.

Proof.

$$\begin{aligned}
\bar{\mathbf{s}}_{k+1} &= \frac{c^p}{(c+k)^p}\mathbf{s}_k + \left(1 - \frac{c^p}{(c+k)^p}\right)\bar{\mathbf{s}}_k \\
&= \frac{c^p}{(c+k)^p}\mathbf{s}_k + \frac{c^p}{(c+k-1)^p}\left(1 - \frac{c^p}{(c+k)^p}\right)\mathbf{s}_{k-1} + \left(1 - \frac{c^p}{(c+k)^p}\right)\left(1 - \frac{c^p}{(c+k-1)^p}\right)\bar{\mathbf{s}}_{k-1} \\
&= \sum_{i=0}^{k}\frac{c^p}{(c+k-i)^p}\prod_{j=0}^{i-1}\left(1 - \frac{c^p}{(c+k-j)^p}\right)\mathbf{s}_{k-i} \\
&\overset{l=k-i}{=} \sum_{l=1}^{k}\underbrace{\frac{c^p}{(c+l)^p}\prod_{j=0}^{k-l-1}\left(1 - \frac{c^p}{(c+k-j)^p}\right)}_{\beta_{k,l}}\mathbf{s}_l.
\end{aligned}$$

If $p = 1$, then

$$\beta_{k,i} = \frac{c}{c+i}\prod_{l=0}^{k-i-1}\frac{k-l}{c+k-l} = \frac{c}{c+i}\frac{k(k-1)(k-2)\cdots(i+1)}{(c+k)(c+k-1)\cdots(c+i+1)} = \frac{c}{c+i}\prod_{j=0}^{c}\frac{i+j+1}{c+k-j}$$

For all p, to show the sum is 1, we do so recursively. At $k = 1$, $\beta_{1,1} = \frac{c^p}{(c+1)^p}$. Now, if $\sum_{i=0}^{k-1}\beta_{k-1,i} = 1$, then for $i \le k-1$

$$\beta_{k,i} = \left(1 - \frac{c^p}{(c+k)^p}\right)\beta_{k-1,i}, \quad i \le k-1$$

and for $i = k$, $\beta_{k,k} = \frac{c^p}{(c+k)^p}$. Then

$$\sum_{i=1}^{k}\beta_{k,i} = \beta_{k,k} + \left(1 - \frac{c^p}{(c+k)^p}\right)\sum_{l=1}^{k-1}\beta_{k-1,i} = \frac{c^p}{(c+k)^p} + \left(1 - \frac{c^p}{(c+k)^p}\right) = 1.$$

□

B Averaging

In the vanilla Frank-Wolfe method, we have two players s and x, and as $x \to x^*$, s may oscillate around the solution fascet however it would like, so that its average is x^* but $\|s - x^*\|_2$ remains bounded away from 0. However, we now show that if we replace s with $\bar{s}$, whose velocity slows down, then it must be that $\|s - x^*\|_2$ decays.

Lemma B.1 (Continuous averaging)**.** *Consider some vector trajectory $v(t) \in \mathbb{R}^n$, and suppose*

- $\|v(t)\|_2 \le D$ *for arbitrarily large* t
- $\|v'(t)\|_2 = \beta(t)D$
- $\frac{1}{2}\|\int_t^\infty \gamma(\tau)v(\tau)d\tau\|_2^2 = O(1/t^q)$ *for* $q > 0$.

Then $\|v(t)\|_2^2 \leq O(t^{q/2+p-1})$.

Proof. We start with the orbiting property.

$$\frac{d}{dt}\left(\frac{1}{2}\|\int_t^\infty \gamma(\tau)v(\tau)d\tau\|_2^2\right) = -\int_t^\infty \gamma(t)\gamma(\tau)v(\tau)^T v(t)d\tau \leq 0.$$

Since this is happening asymptotically, then the negative derivative of the LHS must be upper bounded by the negative derivative of the RHS. That is, if a function is decreasing asymptotically at a certain rate, then its negative derivative should be decaying asymptotically at the negative derivative of this rate. So,

$$\int_t^\infty \gamma(\tau)v(\tau)^T v(t)d\tau \leq O(1/t^q).$$

This indicates either that $\|v(t)\|_2$ is getting smaller (converging) or $v(t)$ and its average are becoming more and more uncorrelated (orbiting).

Doing the same trick again with the negative derivative,

$$-\frac{d}{dt}\int_t^\infty \gamma(\tau)v(\tau)^T v(t)d\tau = \gamma(t)\|v(t)\|_2^2 - \int_t^\infty \gamma(\tau)v(\tau)^T v'(t)d\tau$$

By similar logic, this guy should also be decaying at a rate $O(1/t^{q+1})$, so

$$\gamma(t)\|v(t)\|_2^2 \leq \frac{C_2}{t^{q+1}} + \int_t^\infty \gamma(\tau)v(\tau)^T v'(t)d\tau \leq \frac{C_2}{t^{q+1}} + \underbrace{\|\int_t^\infty \gamma(\tau)v(\tau)d\tau\|_2}_{\leq O(1/t^{q/2})} D\beta(t) = \frac{C_2}{t^{q+1}} + \frac{C_3}{t^{q/2+p}}$$

Therefore

$$\|v(t)\|_2^2 \leq \frac{C_2}{t^q} + \frac{C_3}{t^{q/2+p-1}} = O(\frac{1}{t^{q/2+p-1}}).$$

□

Corollary B.2. *Suppose* f *is* μ*-strongly convex. Then*

$$\|\bar{s}(t) - x(t)\|_2^2 \leq Ct^{-(q/2+p-1)}$$

for some constant $C > 0$*.*

Proof. Taking $v(t) = \bar{s}(t) - x(t)$, it is clear that if $\beta(t) \geq \gamma(t)$ then the first two conditions are satisfied. In the third condition, note that

$$\int_t^\infty \gamma(\tau)(\bar{s}(\tau) - x(\tau))d\tau = \int_t^\infty \dot{x}(\tau)d\tau = x^* - x(t)$$

and therefore

$$\frac{1}{2}\|\int_t^\infty \gamma(\tau)v(\tau)d\tau\|_2^2 = \frac{1}{2}\|x^* - x(t)\|_2^2 \leq \mu(f(x) - f^*)$$

by strong convexity. □

Lemma B.3 (Discrete averaging)**.** *Consider some vector trajectory* $\mathbf{v}_k \in \mathbb{R}^n$*. Then the following properties cannot all be true.*

- $\|\mathbf{v}_k\|_2 \leq D$ *for arbitrarily large* k
- $\|\mathbf{v}_{k+1} - \mathbf{v}_k\|_2 \leq \beta_k D$
- $\frac{1}{2}\|\sum_{i=k}^\infty \gamma_k \mathbf{v}_k\|_2^2 \leq \frac{C_1}{k^q}$ *for* $q > 0$*.*

Then $\|\mathbf{v}_k\|_2^2 \leq O(k^{q/2+p-1})$.

Proof. The idea is to recreate the same proof steps as in the previous lemma. Note that the claimis not that these inequalities happen at each step, but that they must hold asymptotically in order for the asymptotic decay rates to hold. So

$$\frac{1}{2}\|\sum_{i=k}^{\infty}\gamma_i\mathbf{v}_i\|_2^2-\frac{1}{2}\|\sum_{i=k}^{\infty}\gamma_{i+1}\mathbf{v}_{i+1}\|_2^2=\frac{\gamma_k^2}{2}\|\mathbf{v}_k\|_2^2+\gamma_k\mathbf{v}_k^T\left(\sum_{i=k}^{\infty}\gamma_{i+1}\mathbf{v}_{i+1}\right)\leq\frac{C_1}{(k+1)^q}-\frac{C_1}{k^q}=\frac{C_2}{k^q}$$

and therefore

$$\frac{\gamma_k}{2}\|\mathbf{v}_k\|_2^2+\mathbf{v}_k^T\left(\sum_{i=k}^{\infty}\gamma_{i+1}\mathbf{v}_{i+1}\right)\leq\frac{C_1}{(k+1)^{q+1}}-\frac{C_1}{k^{q+1}}=\frac{C_2}{k^{q+1}}.$$

Next,

$$\begin{aligned}&\mathbf{v}_k^T\left(\sum_{i=k}^{\infty}\gamma_{i+1}\mathbf{v}_{i+1}\right)-\mathbf{v}_{k+1}^T\left(\sum_{i=k+1}^{\infty}\gamma_{i+1}\mathbf{v}_{i+1}\right)+\mathbf{v}_k^T\left(\sum_{i=k+1}^{\infty}\gamma_{i+1}\mathbf{v}_{i+1}\right)-\mathbf{v}_k^T\left(\sum_{i=k+1}^{\infty}\gamma_{i+1}\mathbf{v}_{i+1}\right)\\ =\;&\gamma_{k+1}\mathbf{v}_k^T\mathbf{v}_{k+1}+(\mathbf{v}_k-\mathbf{v}_{k+1})^T\left(\sum_{i=k+1}^{\infty}\gamma_{i+1}\mathbf{v}_{i+1}\right)\end{aligned}$$

$$\frac{\gamma_k}{2}\|\mathbf{v}_k\|_2^2-\frac{\gamma_{k+1}}{2}\|\mathbf{v}_{k+1}\|_2^2+\mathbf{v}_k^T\left(\sum_{i=k}^{\infty}\gamma_{i+1}\mathbf{v}_{i+1}\right)-\mathbf{v}_{k+1}^T\left(\sum_{i=k+1}^{\infty}\gamma_{i+1}\mathbf{v}_{i+1}\right)=$$

$$\frac{\gamma_k}{2}\|\mathbf{v}_k\|_2^2\underbrace{-\frac{\gamma_{k+1}}{2}\|\mathbf{v}_{k+1}\|_2^2+\gamma_{k+1}\mathbf{v}_k^T\mathbf{v}_{k+1}}_{-\frac{\gamma_{k+1}}{2}\|\mathbf{v}_{k+1}-\mathbf{v}_k\|_2^2+\frac{\gamma_{k+1}}{2}\|\mathbf{v}_k\|_2^2}+(\mathbf{v}_k-\mathbf{v}_{k+1})^T\left(\sum_{i=k+1}^{\infty}\gamma_{i+1}\mathbf{v}_{i+1}\right)\leq\frac{C_3}{k^{q+1}}$$

Therefore

$$\frac{\gamma_k+\gamma_{k+1}}{2}\|\mathbf{v}_k\|_2^2\leq\frac{C_3}{k^{q+1}}+\underbrace{(\mathbf{v}_{k+1}-\mathbf{v}_k)^T\left(\sum_{i=k+1}^{\infty}\gamma_{i+1}\mathbf{v}_{i+1}\right)}_{O(\beta_k/k^{q/2})}+\underbrace{\frac{\gamma_{k+1}}{2}\|\mathbf{v}_{k+1}-\mathbf{v}_k\|_2^2}_{O(\gamma_k\beta_k^2)}$$

Finally,

$$\|\mathbf{v}_k\|_2^2\leq\frac{C_3}{k^q}+\frac{C_4}{k^{q/2+p-1}}+\frac{C_5}{k^{2p}}=O(1/k^{\min\{q/2+p-1,2p\}}).$$

□

Corollary B.4. *Suppose f is μ-strongly convex. Then if $f(x)-f^*=O(k^{-q})$*

$$\|\bar{\mathbf{s}}_k-\mathbf{x}_k\|_2^2\leq C\max\{k^{-(q/2+p-1)},k^{-2p}\}$$

for some constant $C>0$.

Proof. Taking $\mathbf{v}_k=\bar{\mathbf{s}}_k-\mathbf{x}_k$, it is clear that if $\beta(t)\geq\gamma(t)$ then the first two conditions are satisfied. In the third condition, note that

$$\sum_{i=k}^{\infty}\gamma_i(\bar{\mathbf{s}}_i-\mathbf{x}_i)=\sum_{i=k}^{\infty}\mathbf{x}_{i+1}-\mathbf{x}_i=\mathbf{x}^*-\mathbf{x}_k$$

and therefore

$$\frac{1}{2}\|\sum_{i=k}^{\infty}\gamma_i(\bar{\mathbf{s}}_i-\mathbf{x}_i)\|_2^2=\frac{1}{2}\|\mathbf{x}^*-\mathbf{x}_k\|_2^2\leq\mu(f(\mathbf{x}_k)-f^*)$$

by strong convexity. □

C Global rates

Lemma C.1 (Continuous energy function decay)**.** *Suppose $c \geq q-1$, and*

$$g(t) \leq -\int_0^t \frac{\exp(\alpha(\tau))}{\exp(\alpha(t))}\frac{(c+\tau)^c}{(c+t)^c}\frac{C_1}{(b+\tau)^r}d\tau$$

Then

$$g(t) \leq -\frac{\left(1-\frac{\alpha(1)}{\exp(\alpha(1))}\right)C_1}{\alpha(t)(1-p)}(c+t)^{1-r}$$

Proof.

$$\begin{aligned}
\frac{g(t)}{C_1}\exp(\alpha(t))(c+t)^c &\leq -\int_0^t \exp(\alpha(\tau))(c+\tau)^{c-r}d\tau \\
&= -\int_0^t \sum_{k=1}^{\infty}\frac{\alpha(\tau)^k}{k!}(c+\tau)^{c-r}d\tau \\
&= -\int_0^t \sum_{k=1}^{\infty}\frac{c^{kp}}{k!(1-p)^k}(c+\tau)^{k-pk+c-r}d\tau \\
&\overset{\text{Fubini}}{=} -\sum_{k=1}^{\infty}\frac{c^{kp}}{(1-p)^k k!}\int_0^t (c+\tau)^{k-pk+c-r}d\tau \\
&= -\sum_{k=1}^{\infty}\frac{c^{kp}}{(1-p)^k k!}\frac{(c+t)^{k-pk+c-r+1}-c^{k-pk+c-r+1}}{k-pk+c-r+1} \\
&= -\sum_{k=1}^{\infty}\frac{1}{(k+1)!}\left(\frac{c^p}{(c+t)^{p-1}(1-p)}\right)^k(c+t)^{1+c-r}\underbrace{\frac{1}{(1-p)+(c-r+1)/k}\frac{k+1}{k}}_{\geq C_2} \\
&\leq -C_2\sum_{k=1}^{\infty}\frac{(c+t)^{1+c-r}}{(k+1)!}\frac{\alpha(t)^{k+1}}{\alpha(t)} \\
&= -\frac{C_2(c+t)^{1+c-r}}{\alpha(t)}(\exp(\alpha(t))-\alpha(1))
\end{aligned}$$

Then

$$\begin{aligned}
g(t) &\leq -\frac{C_1C_2}{\alpha(t)}(c+t)^{1-r}\left(1-\frac{\alpha(1)}{\exp(\alpha(t))}\right) \\
&\leq -\frac{C_1C_2}{\alpha(t)}(c+t)^{1-r}\left(1-\frac{\alpha(1)}{\exp(\alpha(1))}\right) \\
&\leq -\frac{C_1C_3}{\alpha(t)}(c+t)^{1-r}
\end{aligned}$$

where $C_3 = C_2\left(1-\frac{\alpha(1)}{\exp(\alpha(1))}\right)$ and $C_2 = \frac{1}{1-p}$ satisfies the condition. □

Theorem C.2 (Continuous global rate)**.** *Suppose $0 < p < 1$. Then the averaged FW flow decays as $O(1/t^{1-p})$.*

Proof. Consider the error function

$$g(t) := \nabla f(x(t))^T(\bar{s}(t)-x(t)), \qquad g(0) = 0.$$

$$
\begin{aligned}
\dot{g}(t) &= \frac{\partial}{\partial t}\nabla f(x)^T(\bar{s}-x) \\
&= \left(\frac{\partial}{\partial t}\nabla f(x)^T\right)(\bar{s}-x) + \nabla f(x)^T\left(\frac{\partial}{\partial t}(\bar{s}-x)\right) \\
&= \underbrace{\left(\frac{\partial}{\partial t}\nabla f(x)^T\right)}_{=\dot{x}^T\nabla^2 f(x)}(\bar{s}-x) + \nabla f(x)^T\left(\beta(t)(s(t)-\bar{s}(t)) - \gamma(t)(\bar{s}(t)-x(t))\right) \\
&\leq \gamma\underbrace{(\bar{s}-x)^T\nabla^2 f(x)(\bar{s}-x)}_{\leq 4LD^2\gamma(t)} + \underbrace{\beta(t)\nabla f(x)^T(s(t)-x(t))}_{-\beta(t)\mathbf{gap}(x)} - (\beta(t)+\gamma(t))\underbrace{\nabla f(x)^T(\bar{s}(t)-x(t))}_{=g(t)}
\end{aligned}
$$

$$
\begin{aligned}
g(t) &\leq \int_0^t \frac{\exp(\alpha(\tau))}{\exp(\alpha(t))}\frac{(c+\tau)^c}{(c+t)^c}\underbrace{\left(\frac{4LD^2c}{c+\tau} - \frac{b^p}{(b+\tau)^p}\mathbf{gap}(x(\tau))\right)}_{A(\tau)}d\tau \\
\dot{h}(t) &\leq \gamma(t)g(t) \leq \gamma(t)\int_0^t \underbrace{\frac{\exp(\alpha(\tau))}{\exp(\alpha(t))}\frac{(c+\tau)^c}{(c+t)^c}}_{\mu(\tau)}A(\tau)d\tau
\end{aligned}
$$

In order for $h(t)$ to decrease, it must be that $\dot{h}(t) \leq 0$. However, since $\mu(\tau) \geq 0$ for all $\tau \geq 0$, it must be that $A(\tau) \leq 0$, e.g.

$$
\frac{c^p}{(c+\tau)^p}\mathbf{gap}(x(\tau)) \geq \frac{4LD^2c}{c+\tau}.
$$

which would imply $h(t) = O(1/(c+t)^{1-p})$. Let us therefore test the candidate solution

$$
h(t) = \frac{C_3}{(c+t)^{1-p}}.
$$

Additionally, from Lemma C.1, if

$$
A(\tau) \leq -\frac{C_1}{(c+t)} \quad \Rightarrow \quad g(t) \leq -\frac{C_1}{\alpha(t)(1-p)}
$$

and therefore

$$
\begin{aligned}
\dot{h}(t) &\leq \gamma(t)g(t) \leq -\frac{c}{c+t}\frac{C_1}{c^p}\cdot(c+t)^{p-1} \\
\int_0^t \dot{(}h(\tau))d\tau &\leq \frac{C_1 c}{(1-p)c^p}(c+t)^{p-1}
\end{aligned}
$$

which satisfies our candidate solution for $C_3 = \frac{C_1 c}{(p-1)c^p}$. □

This term $(\bar{s}-x)^T\nabla^2 f(x)(\bar{s}-x) \leq 4LD^2\gamma(t)$ is an important one to consider when talking about local vs global distance. The largest values of the Hessian will probably not correspond to the indices that are "active", and thus this bound is very loose near optimality.

Lemma C.3 (Discrete energy decay)**.** *Suppose* $0 < p < 1$. *Consider the error function*

$$
\mathbf{g}_k := \nabla f(\mathbf{x}_k)^T(\bar{\mathbf{s}}_k - \mathbf{x}_t).
$$

Then

$$
\mathbf{g}_k \leq -\sum_{i=0}^{k-1}\beta_{i,i}(\frac{i+1+c}{k+c})^c\mathbf{gap}(\mathbf{x}_i) - \beta_{k,k}\mathbf{gap}(\mathbf{x}_k) + \frac{4D^2LC_1}{(k+c)^p} + (\frac{c}{k+c})^c\mathbf{g}_0.
$$

where $C_1 = c^p(1 + \frac{1}{(c+1)(1-p)-1})$.

Importantly, C_1 is finite only if $p < 1$. When $p = 1$, the right hand side is at best bounded by a constant, and does not decay, which makes it impossible to show method convergence.

Proof. Define $\mathbf{z}_k = \nabla f(\mathbf{x}_k)$, $\mathbf{g}_k = \mathbf{z}_k^T(\bar{\mathbf{s}}_k - \mathbf{x}_k)$. Then

$$\begin{aligned}
\mathbf{g}_k &= \underbrace{\beta_{k,k}\mathbf{z}_k^T(\mathbf{s}_k - \mathbf{x}_k)}_{-\beta_{k,k}\mathbf{gap}(\mathbf{x}_k)} + \underbrace{\sum_{i=0}^{k-1}\beta_{k,i}\mathbf{z}_k^T(\mathbf{s}_i - \mathbf{x}_k)}_{A} \\
A &= \sum_{i=0}^{k-1} \underbrace{\frac{\beta_{k,i}}{\beta_{k-1,i}}}_{=(1-\frac{c}{(c+i)^p})} \beta_{k-1,i}\mathbf{z}_k^T(\mathbf{s}_i - \mathbf{x}_k) \leq (1 - \frac{c^p}{(c+k)^p})\underbrace{\mathbf{z}_k^T(\bar{\mathbf{s}}_{k-1} - \mathbf{x}_k)}_{=B} \\
B &= \mathbf{z}_k^T(\bar{\mathbf{s}}_{k-1} - \underbrace{(\mathbf{x}_{k-1} + \gamma_{k-1}(\bar{\mathbf{s}}_{k-1} - \mathbf{x}_{k-1}))}_{\mathbf{x}_k}) \\
&= (1-\gamma_{k-1})\mathbf{z}_k^T(\bar{\mathbf{s}}_{k-1} - \mathbf{x}_{k-1}) \\
&= (1-\gamma_{k-1})(\mathbf{z}_k - \mathbf{z}_{k-1})^T\underbrace{(\bar{\mathbf{s}}_{k-1} - \mathbf{x}_{k-1})}_{(\mathbf{x}_k - \mathbf{x}_{k-1})\gamma_{k-1}^{-1}} + (1-\gamma_{k-1})\underbrace{\mathbf{z}_{k-1}^T(\bar{\mathbf{s}}_{k-1} - \mathbf{x}_{k-1})}_{\mathbf{g}_{k-1}} \\
&= \frac{(1-\gamma_{k-1})}{\gamma_{k-1}}\underbrace{(\mathbf{z}_k - \mathbf{z}_{k-1})^T(\mathbf{x}_k - \mathbf{x}_{k-1})}_{\leq L\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_2^2} + (1-\gamma_{k-1})\mathbf{g}_{k-1} \\
&\leq (1-\gamma_{k-1})\gamma_{k-1}L\underbrace{\|\bar{\mathbf{s}}_{k-1} - \mathbf{x}_{k-1}\|_2^2}_{4D^2} + (1-\gamma_{k-1})\mathbf{g}_{k-1}
\end{aligned}$$

Overall,

$$\begin{aligned}
\mathbf{g}_k &\leq -\beta_{k,k}\mathbf{gap}(\mathbf{x}_k) + 4D^2L\gamma_{k-1}(1-\gamma_{k-1})(1-\beta_k) + \underbrace{(1-\beta_k)(1-\gamma_{k-1})}_{\mu_k}\mathbf{g}_{k-1} \\
&= -\beta_{k,k}\mathbf{gap}(\mathbf{x}_k) + 4D^2L\gamma_{k-1}\mu_k + \mu_k\mathbf{g}_{k-1} \\
&= -\beta_{k,k}\mathbf{gap}(\mathbf{x}_k) + 4D^2L\gamma_{k-1}\mu_k - \beta_{k-1,k-1}\mu_k\mathbf{gap}(\mathbf{x}_{k-1}) + 4D^2L\gamma_{k-2}\mu_{k-1}\mu_k + \mu_k\mu_{k-1}\mathbf{g}_{k-2} \\
&= -\sum_{i=0}^{k-1}\beta_{i,i}\mathbf{gap}(\mathbf{x}_i)\prod_{j=i+1}^{k}\mu_j - \beta_{k,k}\mathbf{gap}(\mathbf{x}_k) + 4D^2L\sum_{i=0}^{k}\gamma_{k-i}\prod_{j=i}^{k}\mu_j + \prod_{j=1}^{k}\mu_k\mathbf{g}_0
\end{aligned}$$

Now we compute $\prod_{j=i}^{k}\mu_j$

$$\prod_{j=i}^{k}(1-\gamma_{j-1}) = \prod_{j=i}^{k}\frac{j-1}{c+j-1} = \prod_{j=0}^{c}\frac{i-1+j}{k+j} \leq (\frac{i+c}{k+c})^c$$

Using $1 - \frac{c^p}{(c+k)^p} \leq \exp(-(\frac{c}{c+k})^p)$,

$$\log(\prod_{j=i}^{k}(1 - \frac{c^p}{(c+j)^p})) \leq -\sum_{j=i}^{k}(\frac{c}{c+j})^p \leq -\int_i^k(\frac{c}{c+j})^p dj = \frac{c^p}{p+1}((c+i)^{p+1} - (c+k)^{p+1})$$

and therefore

$$\prod_{j=i}^{k}(1 - \frac{c^p}{(c+k)^p}) \leq \frac{\exp(\frac{c^p}{p+1}(c+i)^{p+1})}{\exp(\frac{c^p}{p+1}(c+k)^{p+1})}$$

which means

$$\prod_{j=i}^{k}\mu_j \leq (\frac{i+c}{k+c})^c\exp(\frac{c^p}{p+1}((c+i)^{p+1} - (c+k)^{p+1})).$$

Now we bound the constant term coefficient.

$$\begin{aligned}
\sum_{i=0}^{k}\gamma_{i-1}\prod_{j=i}^{k}\mu_j &\leq \sum_{i=0}^{k}\underbrace{\frac{c}{(c+i-1)}(\frac{i+c}{k+c})^c}_{\text{max at } i=0}\underbrace{\exp(\frac{c^p}{p+1}((c+i)^{p+1}-(c+k)^{p+1}))}_{\leq\frac{1}{(k-i)^{(c+1)(1-p)}}} \\
&\overset{C-S}{\leq} \frac{c}{c-1}\frac{c^c}{(k+c)^c}\sum_{i=0}^{k-1}\frac{1}{(k-i)^{(c+1)(1-p)}} \\
&\leq \frac{c}{c-1}\frac{c^c}{(k+c)^c}\frac{1}{(c+1)(1-p)-1}\frac{1}{(1-k)^{(c+1)(1-p)-1}} \\
&\leq \frac{C_1}{(k+c)^p}\frac{1}{(1-k)^{(c+1)(1-p)-1}}
\end{aligned}$$

where $(*)$ if c is chosen such that $(c+1)(1-p)>1$ and $C_1>0$ big enough. Note that necessarily, $p<1$, and the size of C_1 depends on how close p is to 1.

Also, to simplify terms,

$$\prod_{j=i}^{k}\mu_j \leq (\frac{i+c}{k+c})^c\underbrace{\exp(\frac{c^p}{p+1}((c+i)^{p+1}-(c+k)^{p+1}))}_{\leq 1}.$$

Now, we can say

$$\mathbf{g}_k \leq -\sum_{i=0}^{k-1}\beta_{i,i}(\frac{i+1+c}{k+c})^c\mathbf{gap}(\mathbf{x}_i)-\beta_{k,k}\mathbf{gap}(\mathbf{x}_k)+\frac{4D^2LC_1}{(k+c)^p}+(\frac{c}{k+c})^c\mathbf{g}_0.$$

□

Theorem C.4 (Global rate, $p<1$.)**.** *Suppose* $0<p<1$ *and* $c\geq\frac{c-1}{c^p}$. *Then* $h(\mathbf{x}_k)=:\mathbf{h}_k=O(\frac{1}{(k+c)^p})$.

Proof. Start with

$$\mathbf{g}_k \leq -\sum_{i=0}^{k-1}\beta_{i,i}(\frac{i+1+c}{k+c})^c\mathbf{gap}(\mathbf{x}_i)-\beta_{k,k}\mathbf{gap}(\mathbf{x}_k)+\frac{4D^2LC_1}{(k+c)^p}+(\frac{c}{k+c})^c\mathbf{g}_0.$$

$$\begin{aligned}
\mathbf{h}_{k+1}-\mathbf{h}_k &\leq \gamma_k\mathbf{g}_k+2\gamma_kLD^2 \\
&\leq -\frac{c}{c+k}\frac{1}{(k+c)^c}\sum_{i=0}^{k-1}\beta_{i,i}(i+c+1)^c\mathbf{h}_i-\frac{c}{c+k}\beta_{k,k}\mathbf{h}_k \\
&\quad +2D^2L\gamma_k(\frac{2C_1}{(k+c)^p}+\gamma_k)+\frac{c}{c+k}\frac{c^c}{(k+c)^c}\mathbf{g}_0 \\
&\leq -\frac{c^{p+1}}{c+k}\frac{1}{(k+c)^c}\sum_{i=0}^{k-1}(i+c)^{c-p}\mathbf{h}_i-\frac{c}{c+k}\beta_{k,k}\mathbf{h}_k \\
&\quad +2D^2L\gamma_k(\frac{2C_1}{(k+c)^p}+\gamma_k)+\frac{c}{c+k}\frac{c^c}{(k+c)^c}\mathbf{g}_0
\end{aligned}$$

Suppose $\mathbf{h}_k \leq \frac{C_2}{(k+c)^p}$. Then

$$\begin{aligned}
\mathbf{h}_{k+1} - \mathbf{h}_k &\leq \underbrace{-\frac{c^{p+1}}{c+k}\frac{C_2}{(k+c)^c}\sum_{i=0}^{k-1}(c+i)^{c-2p} - \frac{cC_2}{c+k}\frac{c^p}{(c+k)^{2p}}}_{-AC_2} \\
&\quad \underbrace{+2D^2L\frac{c}{c+k}\left(\frac{2C_1}{(k+c)^p} + \frac{c}{c+k}\right) + \frac{c}{c+k}\frac{c^c}{(k+c)^c}\mathbf{g}_0}_{B} \\
B &= \frac{c}{c+k}\left(2D^2L\left(\frac{2C_1}{(k+c)^p} + \frac{c}{c+k}\right) + \frac{c^c}{(k+c)^c}\mathbf{g}_0\right) \\
&\leq \frac{2cD^2L(2C_1+c) + c^{c+1}\mathbf{g}_0}{(c+k)^p(c+k)} \\
&=: \frac{C_3}{(c+k)^{p+1}}
\end{aligned}$$

where $c > 1$. Then,

$$\begin{aligned}
(k+c)^c A &= \frac{c^{p+1}}{c+k}\sum_{i=0}^{k-1}(c+i)^{c-2p} + \frac{c}{c+k}\frac{c^p}{(c+k)^{2p-1}} \\
&\geq \frac{c^{p+1}}{c+k}\frac{(c+k-1)^{c-2p+1} - (c+1)^{c-2p+1}}{c-2p+1} + \frac{c}{c+k}\frac{c^p}{(c+k)^{2p-1}} \\
&= \frac{c^{p+1}}{c-2p+1}\frac{(c+k-1)^{c-2p+1}}{c+k} + O(1/k) \\
&\overset{c\geq 2p+1}{\geq} \frac{c^{p+1}}{c-2p+1} + O(1/k) \\
&\overset{k \text{ big enough}}{\geq} \frac{c^{p+1}}{c-1}
\end{aligned}$$

Therefore,

$$\mathbf{h}_{k+1} \leq \frac{C_2(1-\frac{c^{p+1}}{c-1})}{(k+c)^p} + \frac{C_3}{(c+k)^{p+1}}.$$

Define $\epsilon = \frac{2c^{p+1}}{c^{p+1}+1-c}$ and pick $C_2 > \frac{C_3}{c\epsilon}$. By assumption, $\epsilon > 0$. Consider $k > K$ such that for all k,

$$\frac{(k+c+1)^p}{(k+c)^p} \leq 1 + \frac{\epsilon}{2}, \qquad \frac{C_3}{c+k} \leq \frac{\epsilon}{2}\frac{(c+k)^p}{(c+k+1)^p}.$$

Then

$$\mathbf{h}_{k+1} \leq \frac{C_2(1-\frac{\epsilon}{2})}{(k+c+1)^p} + \frac{\frac{\epsilon}{2}}{(k+c+1)^p} \leq \frac{C_2}{(k+c+1)^p}.$$

We have now proved the inductive step. Picking $C_2 \geq \mathbf{h}_0$ gives the starting condition, completing the proof.

□

D Local rates

Lemma D.1 (Local convergence)**.** *Define, for all t,*

$$\tilde{s}(t) = \operatorname*{argmin}_{\tilde{s}\in\mathbf{conv}(\mathcal{S}(\mathbf{x}^*))} \|s(t) - \tilde{s}\|_2, \qquad \hat{s}(t) = \bar{\beta}_t^{-1}\int_0^t \beta_{t,\tau}\tilde{s}(\tau)d\tau. \tag{6}$$

e.g., $\hat{s}(t)$ is the closest point in the convex hull of the support of $\mathbf{x}^*$ to the the point $s(t) = \mathrm{LMO}_{\mathcal{D}}(x(t))$.

Then,

$$\|\bar{s}(t) - \hat{s}(t)\|_2 \leq \frac{c_1}{(c+t)^c}.$$

The proof of this theorem actually does not really depend on how well the FW method works inherently, but rather is a consequence of the averaging. Intuitively, the proof states that after the manifold has been identified, all new support components must also be in the convex hull of the support set of the *optimal* solution; thus, in fact $s_2(t) - \hat{s}(t) = 0$. However, because the accumulation term in the flow actually is not a true average until $t \to +\infty$, there is a pesky normalization term which must be accounted for. Note, importantly, that this normalization term does not appear in the method, where the accumulation weights always equal 1 (pure averaging).

Proof. First, note that

$$\bar{s}(t) - \hat{s}(t) = \bar{s}(t) - \bar{\beta}_t^{-1} \int_0^t \beta_{t,\tau} \tilde{s}(\tau) d\tau = \int_0^t \beta_{t,\tau}(s(\tau) - \bar{\beta}_t^{-1}\tilde{s}(\tau)) d\tau.$$

Using triangle inequality,

$$\|\bar{s}(t) - \hat{s}(t)\|_2 \leq \underbrace{\| \int_0^{\bar{t}} \beta_{t,\tau}(s(\tau) - \bar{\beta}_t^{-1}\tilde{s}(\tau)) d\tau \|_2}_{\epsilon_1} + \underbrace{\| \int_{\bar{t}}^t \beta_{t,\tau}(s(\tau) - \bar{\beta}_t^{-1}\tilde{s}(\tau)) d\tau \|_2}_{\epsilon_2}.$$

Expanding the first term, via Cauchy Scwhartz for integrals, we can write, elementwise,

$$\int_0^{\bar{t}} \beta_{t,\tau}(s(\tau)_i - \bar{\beta}_t^{-1}\tilde{s}(\tau)_i) d\tau \leq \int_0^{\bar{t}} \beta_{t,\tau} d\tau \int_0^{\bar{t}} |s(\tau)_i - \bar{\beta}_t^{-1}\tilde{s}(\tau)_i| d\tau$$

and thus,

$$\| \int_0^t \beta_{t,\tau}(s(\tau) - \bar{\beta}_t^{-1}\tilde{s}(\tau)) d\tau \| \leq \int_0^{\bar{t}} \beta_{t,\tau} d\tau \underbrace{\| \int_0^t s(\tau)_i - \bar{\beta}_t^{-1}\tilde{s}(\tau)_i d\tau \|_2}_{\leq 2D(1+\bar{\beta}_t^{-1})\bar{t}}.$$

and moreover,

$$\int_0^{\bar{t}} \beta_{t,\tau} d\tau = \int_0^{\bar{t}} \frac{c(c+\tau)^{c-1}}{(c+t)^c} d\tau = \frac{\hat{c}_0}{(c+t)^c}$$

since $\int_0^{\bar{t}} b(b+\tau)^{b-1} d\tau$ does not depend on t. Thus the first error term

$$\epsilon_1 \leq \frac{2\hat{c}_0 D\bar{t}(1+\bar{\beta}_t^{-1})}{(c+t)^c} \leq \frac{c_0}{(c+t)^c}$$

where

$$\hat{c}_0 := 2D\bar{t} \int_0^{\bar{t}} c(c+\tau)^{c-1} d\tau.$$

In the second error term, *because the manifold has now been identified*, $s(\tau) = \tilde{s}(\tau)$, and so

$$\int_{\bar{t}}^t \beta_{t,\tau}(s(\tau) - \bar{\beta}_t^{-1}\tilde{s}(\tau)) d\tau = \int_{\bar{t}}^t \beta_{t,\tau}(1 - \bar{\beta}_t^{-1}) s(\tau) d\tau$$

and using the same Cauchy-Schwartz argument,

$$\int_{\bar{t}}^t \beta_{t,\tau}(1 - \bar{\beta}_t^{-1}) s(\tau) d\tau \leq D \int_{\bar{t}}^t \beta_{t,\tau}(1 - \bar{\beta}_t^{-1}) d\tau.$$

The term

$$1 - \bar{\beta}_t^{-1} = |1 - \frac{1}{1 - (\frac{c}{c+t})^c}| = \frac{c^c}{(c+t)^c - c^c} \leq \frac{2c^2}{(c+t)^c}$$

and thus

$$\epsilon_t \leq \int_{\bar{t}}^{t} \beta_{t,\tau}(1-\bar{\beta}_t^{-1})d\tau \leq \frac{2c^3}{(c+t)^{2c}} \int_{\bar{t}}^{t} (c+\tau)^{c-1} = \frac{2c^2}{(c+t)^{2c}}((c+t)^c - (c+\bar{t})^c) \leq \frac{2c^2}{(c+t)^c}.$$

Thus,

$$\|\bar{s}(t) - \hat{s}(t)\|_2 \leq \frac{\hat{c}_0 + 2c^2}{(c+t)^c} = O(\frac{1}{(c+t)^c}).$$

□

Corollary D.2 (Local flow rate)**.** *Suppose that for all $x \in \mathcal{D}$, $\|\nabla f(x)\|_2 \leq G$ for some G large enough. Consider $\gamma(t) = \beta(t) = \frac{c}{c+t}$. Then the ODE*

$$\dot{h}(x(t)) = \gamma(t)\nabla f(x)^T(\bar{s} - x)$$

has solutions $h(t) = O(\frac{\log(t)}{(c+t)^c})$ when $t \geq \bar{t}$.

Proof. First, we rewrite the ODE in a more familiar way, with an extra error term

$$\dot{h}(x(t)) = \gamma(t)\nabla f(x)^T(\bar{s} - \hat{s}) + \gamma(t)\nabla f(x)^T(\hat{s} - x)$$

where $\hat{s}$ is as defined in (6). By construction, $\hat{s}$ is a convex combination of $\tilde{s} \in \mathcal{S}(\mathbf{x}^*)$. Moreover, after $t \geq \bar{t}$, $\mathcal{S}(\bar{x}(t)) = \mathcal{S}(\mathbf{x}^*)$, and thus

$$\nabla f(x)^T(\hat{s}(t) - x) = \nabla f(x)^T(s(t) - x) = -\mathbf{gap}(t) \leq -h(t).$$

Then, using Cauchy-Schwartz, and piecing it together,

$$h(t) = \nabla f(x)^T(\hat{s}(t) - x) \leq G\gamma(t)\|\bar{s} - \hat{s}\|_2 - \gamma(t)h(t) \leq \frac{G\gamma(t)c_1}{(c+t)^c} - \gamma(t)h(t).$$

Let us therefore consider the system

$$\dot{h}(x(t)) = \frac{2GD\gamma(t)}{(c+t)^c} - \gamma(t)h(x(t)).$$

The solution to this ODE is

$$h(t) = \frac{h(0)c^c + 2GDc\log(c+t) - 2GDc\log(c)}{(c+t)^c} = O(\frac{\log(t)}{(c+t)^c}).$$

□

Lemma D.3 (Local averaging error)**.** *Define, for all k,*

$$\tilde{\mathbf{s}}_k = \operatorname*{argmin}_{\tilde{\mathbf{s}} \in \mathbf{conv}(\mathcal{S}(\mathbf{x}^*))} \|\mathbf{s}_k - \tilde{\mathbf{s}}\|_2, \qquad \hat{\mathbf{s}}_k = \bar{\beta}_k^{-1} \sum_{i=1}^{k} \beta_{k,i}\tilde{\mathbf{s}}_i.$$

e.g., $\tilde{\mathbf{s}}(k)$ is the closest point in the convex hull of the support of $\mathbf{x}^*$ to the the point $\mathbf{s}_k = \mathrm{LMO}_{\mathcal{D}}(\mathbf{x}_k)$. Then,

$$\|\bar{\mathbf{s}}_k - \hat{\mathbf{s}}_k\|_2 \leq \frac{c_2}{k^c}.$$

Proof. First, note that

$$\bar{\mathbf{s}}_k - \hat{\mathbf{s}}_k = \bar{\mathbf{s}}_k - \sum_{i=1}^{k} \beta_{k,i}\tilde{\mathbf{s}}_i = \sum_{i=1}^{k} \beta_{k,i}(\mathbf{s}_i - \tilde{\mathbf{s}}_i).$$

Using triangle inequality,

$$\|\bar{\mathbf{s}}_k - \hat{\mathbf{s}}_k\|_2 \leq \underbrace{\|\sum_{i=1}^{\bar{k}} \beta_{k,i}(\mathbf{s}_i - \tilde{\mathbf{s}}_i))\|_2}_{\epsilon} + \underbrace{\|\sum_{i=\bar{k}}^{k} \beta_{k,i}(\mathbf{s}_i - \tilde{\mathbf{s}}_i)\|_2}_{0}.$$

where the second error term is 0 since the manifold has been identified, so $\tilde{\mathbf{s}}_i = \mathbf{s}_i$ for all $i \geq \bar{k}$.

Expanding the first term, using a Holder norm (1 and ∞ norm) argument,

$$\begin{aligned}
\|\sum_{i=1}^{\bar{k}} \beta_{k,i}(\mathbf{s}_i - \tilde{\mathbf{s}}_i)\|_2 &\leq 2D \sum_{i=1}^{\bar{k}} \beta_{k,i} \\
&= 2D \sum_{i=1}^{\bar{k}} \frac{c}{(c+i)} \prod_{j=0}^{k-i-1} (1 - \frac{c}{(c+k-j)}) \\
&= 2D \sum_{i=1}^{\bar{k}} \frac{c}{(c+i)} \prod_{j=0}^{c} \frac{i-1+j}{c+k-j} \\
&\leq 2D (\frac{\bar{k}-1+c}{k})^c \sum_{i=1}^{\bar{k}} \frac{c}{(c+i)} = O(1/k^c).
\end{aligned}$$

□

Corollary D.4 (Local convergence rate bounds)**.** *Suppose that for all $\mathbf{x} \in \mathcal{D}$, $\|\nabla f(\mathbf{x})\|_2 \leq G$ for some G large enough. Define also r the decay constant of $\|\bar{\mathbf{s}}_k - \mathbf{x}_k\|_2^2$ $(= O(1/k^r))$. Consider $\gamma_k = \beta_k = \frac{c}{c+k}$. Then the difference equation*

$$\mathbf{h}(\mathbf{x}_{k+1}) - \mathbf{h}(\mathbf{x}_k) \leq \gamma_k \nabla f(\mathbf{x})^T (\bar{\mathbf{s}}_k - \mathbf{x}_k) + \frac{C}{k^r}$$

is satisfied with candidate solution $\mathbf{h}(\mathbf{x}_k) = C_4 \max\{\frac{\log(k)}{(c+k)^c}, \frac{1}{k^{r+1}}\}$ when $k \geq \bar{k}$.

Proof. First, we rewrite the ODE in a more familiar way, with an extra error term

$$\mathbf{h}(\mathbf{x}_{k+1}) - \mathbf{h}(\mathbf{x}_k) = \gamma_k \underbrace{\nabla f(\mathbf{x}_k)^T (\bar{\mathbf{s}}_k - \hat{\mathbf{s}}_k)}_{\leq \gamma_k G \|\bar{\mathbf{s}}_k - \hat{\mathbf{s}}_k\|_2} + \gamma_k \nabla f(\mathbf{x}_k)^T (\hat{\mathbf{s}}_k - \mathbf{x}_k) + \frac{C}{k^{r+2}}$$

where $\hat{\mathbf{s}}_k$ is as defined in (6). By construction, $\hat{\mathbf{s}}_k$ is a convex combination of $\tilde{\mathbf{s}}_i \in \mathcal{S}(\mathbf{x}^*)$. Moreover, after $k \geq \bar{k}$, $\mathcal{S}(\bar{\mathbf{x}}_k) = \mathcal{S}(\mathbf{x}^*)$, and thus

$$\nabla f(\mathbf{x}_k)^T (\hat{\mathbf{s}}_k - \mathbf{x}_k) = \nabla f(\mathbf{x}_k)^T (\mathbf{s}_k - \mathbf{x}_k) = -\mathbf{gap}(\mathbf{x}_k) \leq -\mathbf{h}(\mathbf{x}_k).$$

Then, piecing it together,

$$\mathbf{h}(\mathbf{x}_{k+1}) - \mathbf{h}(\mathbf{x}_k) \leq \gamma_k G \|\bar{\mathbf{s}} - \hat{\mathbf{s}}\|_2 - \gamma_k \mathbf{h}(\mathbf{x}_k) + \frac{C}{k^{r+2}} \overset{\text{Lemma D.3}}{\leq} \underbrace{\gamma_k \frac{GC_2}{k^c}}_{\leq C_3/k^{c+1}} - \gamma_k \mathbf{h}(\mathbf{x}_k) + \frac{C}{k^{r+2}}$$

Recursively, we can now show that for $C_4 \geq C + C_3$, if

$$\mathbf{h}(\mathbf{x}_k) \leq C_4 \max\{\frac{\log(k)}{(c+k)^c}, \frac{1}{k^{r+1}}\}$$

then,

$$\begin{aligned}
\mathbf{h}(\mathbf{x}_{k+1}) &\leq \frac{C_3}{k^{c+1}} + \frac{C}{k^{r+1}} + (1 - \gamma_k)\mathbf{h}(\mathbf{x}_k) \\
&\leq \frac{C_3}{k^{c+1}} + \frac{C}{k^{r+1}} + \frac{k}{c+k} C_4 \max\{\frac{\log(k)}{(c+k)^c}, \frac{1}{k^{r+1}}\}.
\end{aligned}$$

If $c \leq r+1$ then

$$\mathbf{h}(\mathbf{x}_{k+1}) \leq \frac{C + C_3}{k^c} + \frac{k}{c+k} \frac{C_4 \log(k)}{(c+k)^c} \leq \frac{k}{c+k} \frac{C_4 \log(k)}{(c+k)^c} \leq \frac{C_4 \log(k+1)}{(c+k+1)^c}$$

for k large enough. Otherwise,

$$\mathbf{h}(\mathbf{x}_{k+1}) \quad \leq \quad \frac{C + C_3}{k^{r+2}} + \frac{k}{c+k}\frac{C_4}{k^{r+1}} \leq \frac{C_4}{(k+1)^{r+1}}.$$

for k large enough. □

Theorem D.5 (Local convergence rate)**.** *Picking $c \geq 3p/2 + 1$, the proposed method* AVGFW *has an overall convergence $\mathbf{h}(\mathbf{x}_k) = O(k^{-3p/2})$.*

Proof. Putting together Theorem C.4, Lemma D.3, and Corollary D.4, we can resolve the constants

$$q = p, \qquad r = \min\{q/2 + p - 1, 2p\} = \frac{3p}{2} - 1$$

and resolves an overall convergence bound of $\mathbf{h}(\mathbf{x}_k) = O(k^{-3p/2})$.

□