6 Conditional Gradient

Goal

Linear optimization, conditional gradient, matrix completion, totally corrective, linear convergence.

Alert 6.1: Convention

Gray boxes are not required hence can be omitted for unenthusiastic readers.
This note is likely to be updated again soon.

Definition 6.2: Problem

We revisit the constrained smooth minimization problem in Lecture 3:

$$\min_{\mathbf{w}\in C} f(\mathbf{w}), \tag{6.1}$$

where $C \subseteq \mathbb{R}^d$ is a closed and bounded convex set. If the objective function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth (see Definition 2.11), we may apply projected gradient Algorithm 3.15. However, our main assumption in this lecture is that the projection operator

$$\mathrm{P}^{\eta}_{\iota_C}(\mathbf{w}) = \mathrm{P}_C(\mathbf{w}) := \operatorname*{argmin}_{\mathbf{z}\in C} \tfrac{1}{2}\|\mathbf{w}-\mathbf{z}\|_2^2$$

is expensive or even hard to compute.

Example 6.3: Matrix completion (e.g. Jaggi and Sulovsky 2010; Zhang et al. 2012)

Suppose we can only observe a few entries in a matrix A, and want to recover the remaining entries, under the assumption that A is low-rank. Naturally, we can formulate this problem as

$$\min_{X:\mathrm{rank}(X)\le k} \sum_{(i,j)\in\mathcal{O}} (A_{ij}-X_{ij})^2,$$

where k (small, fixed) is our guess of the rank of A. Since the rank function is not convex, a lot of work has turned to its convex relaxation:

$$\min_{X:\|X\|_{\mathrm{tr}}\le\lambda} \sum_{(i,j)\in\mathcal{O}} (A_{ij}-X_{ij})^2,$$

where the trace norm $\|X\|_{\mathrm{tr}} := \sum_i \sigma_i(X)$ is the sum of singular values σ_i of X, and $\lambda > 0$ is some fixed hyperparameter. The projection operator onto the norm ball $\mathsf{B}_\lambda := \{X : \|X\|_{\mathrm{tr}} \le \lambda\}$ turns out to have a "closed-form" solution. Let $X = U\Sigma V^\top$ be its singular value decomposition (SVD), then

$$\mathrm{P}_{\mathsf{B}_\lambda}(X) = U\,\mathrm{diag}(\boldsymbol{\gamma})V^\top,$$

where $\boldsymbol{\gamma}$ is the projection of $\boldsymbol{\sigma}$ onto the ℓ_1 norm ball $\{\mathbf{w} : \|\mathbf{w}\|_1 \le \lambda\}$. However, this projection step requires computing the *full* SVD of X, a cubic time operation!

Jaggi, M. and M. Sulovsky (2010). "A Simple Algorithm for Nuclear Norm Regularized Problems". In: *Proceedings of the 27th International Conference on Machine Learning*, pp. 471–478.
Zhang, X., Y. Yu, and D. Schuurmans (2012). "Accelerated Training for Matrix-Norm Regularization: A Boosting Approach". In: *Advances in Neural Information Processing Systems*.

Example 6.4: Structural SVM (e.g. Lacoste-Julien et al. 2013)

Structural SVM is an extension of SVM to the case where the label space can be exponentially large:

$$\min_{\mathbf{w}} \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \frac{1}{n}\sum_{i=1}^{n} \max_{\mathbf{y}\in\mathcal{Y}_i} \left[\Delta(\mathbf{y},\mathbf{y}_i) + \langle \mathbf{w}, \varphi(\mathbf{x}_i,\mathbf{y}) - \varphi(\mathbf{x}_i,\mathbf{y}_i)\rangle\right], \tag{6.2}$$

where $\mathcal{Y}_i$ is the label space for the i-th training sample $(\mathbf{x}_i,\mathbf{y}_i)$, and $\varphi(\mathbf{x},\mathbf{y})$ is the feature transform. In many applications $|\mathcal{Y}_i|$ can be exceedingly large. For our purpose to solve (6.2), it suffices to have the following oracle (a.k.a. loss augmented inference):

$$\operatorname*{argmax}_{\mathbf{y}\in\mathcal{Y}_i} \left[\Delta(\mathbf{y},\mathbf{y}_i) + \langle \mathbf{w}, \varphi(\mathbf{x}_i,\mathbf{y}) - \varphi(\mathbf{x}_i,\mathbf{y}_i)\rangle\right]$$

The objective function of structural SVM is not differentiable, but it is a convex problem. We can of course apply the subgradient algorithm (later), if the above oracle is available (for computing a subgradient).

Note that the structural SVM objective is λ-strongly convex, hence the dual problem is $\frac{1}{\lambda}$-smooth:

$$\begin{aligned} \min_{\boldsymbol{\alpha}\geq 0} \quad & \frac{1}{2\lambda}\|A\boldsymbol{\alpha}\|_2^2 - \langle \mathbf{b}, \boldsymbol{\alpha}\rangle, \\ \text{s.t.} \quad & \forall i, \ \sum_{\mathbf{y}\in\mathcal{Y}_i} \alpha_i(\mathbf{y}) = 1, \end{aligned}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^m$ with $m = \sum_i |\mathcal{Y}_i|$, the columns of A are $\{\frac{1}{n}\varphi(\mathbf{x}_i,\mathbf{y}) - \varphi(\mathbf{x}_i,\mathbf{y}_i) : i = 1,\dots,n, \mathbf{y}\in\mathcal{Y}_i\}$, and the entries of $\mathbf{b}$ are $\{\frac{1}{n}\Delta(\mathbf{y}_i,\mathbf{y}) : i = 1,\dots,n, \mathbf{y}\in\mathcal{Y}_i\}$. In other words, $\boldsymbol{\alpha} \in S := S_1 \times S_2 \times \cdots \times S_n$, a product of n simplexes. Note that $\dim(S_i) = |\mathcal{Y}_i|$ can be exponentially large, so projecting onto S_i can be very expensive.

Further follow-up work on this include Shah et al. (2015), Beck et al. (2015), Wang et al. (2016), and Osokin et al. (2016).

Lacoste-Julien, S., M. Jaggi, M. Schmidt, and P. Pletscher (2013). "Block-Coordinate Frank-Wolfe Optimization for Structured SVMs". In: *Proceedings of the 30th International Conference on Machine Learning*, pp. 53–61.

Shah, N., V. Kolmogorov, and C. H. Lampert (2015). "A multi-plane block-coordinate Frank-Wolfe algorithm for training structural SVMs with a costly max-oracle". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2737–2745.

Beck, A., E. Pauwels, and S. Sabach (2015). "The Cyclic Block Conditional Gradient Method for Convex Optimization Problems". *SIAM Journal on Optimization*, vol. 25, no. 4, pp. 2024–2049.

Wang, Y.-X., V. Sadhanala, W. Dai, W. Neiswanger, S. Sra, and E. Xing (2016). "Parallel and Distributed Block-Coordinate Frank-Wolfe Algorithms". In: *Proceedings of The 33rd International Conference on Machine Learning*, pp. 1548–1557.

Osokin, A., J.-B. Alayrac, I. Lukasewitz, P. Dokania, and S. Lacoste-Julien (2016). "Minding the Gaps for Block Frank-Wolfe Optimization of Structured SVMs". In: *Proceedings of The 33rd International Conference on Machine Learning*, pp. 593–602.

Definition 6.5: Support function

Recall the indicator function of a set C:

$$\iota_C(\mathbf{w}) = \begin{cases} 0, & \mathbf{w}\in C \\ \infty, & \mathbf{w}\notin C \end{cases}.$$

We now define the support function σ_C which is the dual of the indicator function:

$$\sigma(\mathbf{w}^*) = \sup_{\mathbf{w}\in C} \langle \mathbf{w}; \mathbf{w}^*\rangle.$$

The support function is always (closed) convex (even for nonconvex set C). The maximizers in the above definition, collectively denoted as $\partial\sigma(\mathbf{w}^*)$, will be called the subdifferential of σ at $\mathbf{w}^*$. The meaning of subdifferential will become clear in the next lecture.

Remark 6.6: Motivation from optimality condition

Let us motivate the conditional gradient algorithm using the optimality condition (3.8) for (6.1):

$$\begin{aligned}\forall \mathbf{w} \in C, \langle \mathbf{w} - \mathbf{w}_*; \nabla f(\mathbf{w}_*)\rangle \geq 0 &\iff \forall \mathbf{w} \in C, \langle \mathbf{w}_*; -\nabla f(\mathbf{w}_*)\rangle \geq \langle \mathbf{w}; -\nabla f(\mathbf{w}_*)\rangle \\ &\iff \mathbf{w}_* \in (1-\eta)\mathbf{w}_* + \eta \cdot \operatorname*{argmax}_{\mathbf{w}\in C} \langle \mathbf{w}; -\nabla f(\mathbf{w}_*)\rangle \\ &\iff \mathbf{w}_* \in (1-\eta)\mathbf{w}_* + \eta \cdot \partial\sigma(-\nabla f(\mathbf{w}_*)),\end{aligned}$$

leading to the fixed point iteration

$$\mathbf{w}_{t+1} \leftarrow (1-\eta_t)\mathbf{w}_t + \eta_t \cdot \partial\sigma(-\nabla f(\mathbf{w}_t)).$$

Clearly, if $\mathbf{w}_t \in C$ and $\eta_t \in [0,1]$, we automatically have $\mathbf{w}_{t+1} \in C$!

Algorithm 6.7: Conditional gradient (Frank and Wolfe 1956; Dem'yanov and Rubinov 1967)

Algorithm: Conditional gradient (condgrad)

Input: $\mathbf{w}_0 \in C$

1 **for** $t = 0, 1, \ldots$ **do**

2 $\quad \mathbf{z}_t \leftarrow \operatorname*{argmax}_{\mathbf{z}\in C} \langle \mathbf{z}; -\nabla f(\mathbf{w}_t)\rangle$ `// polar operator`

3 $\quad$ choose step size $\eta_t \in [0,1]$

4 $\quad \mathbf{w}_{t+1} \leftarrow (1-\eta_t)\mathbf{w}_t + \eta_t \mathbf{z}_t$ `// convex combination`

Unlike the quadratic proximal step in proximal gradient Algorithm 4.17, the only nontrivial step in line 2 is a linear minimization problem, hence we know $\mathbf{z}_t$ can be chosen as one of the extreme points of C (see Exercise 6.8 below)! Essentially, the algorithm iteratively selects extreme points of C and takes convex combinations of them.

Since $-\nabla f(\mathbf{w}_t)$ is the gradient direction for reducing f, we can also interpret line 2 as finding a direction in C that correlates the most with $-\nabla f(\mathbf{w}_t)$.

Frank, M. and P. Wolfe (1956). "An Algorithm for Quadratic Programming". *Naval Research Logistics Quarterly*, vol. 3, no. 1-2, pp. 95–110.

Dem'yanov, V. F. and A. M. Rubinov (1967). "The Minimization of a Smooth Convex Functional on a Convex Set". *SIAM Journal on Control*, vol. 5, no. 2, pp. 280–294. [English translation of paper in *Vestnik Leningradskogo Universitera, Seriya Matematiki, Mekhaniki i Astronomii* vol. 19, pp. 7–17, 1964].

Exercise 6.8: Convex maximization returns extreme points

Recall that $\mathbf{w} \in C$ is an extreme point (of C) if it does not lie on the line segment of any two points in C. In other words, if $\mathbf{w} \in [\mathbf{w}_1, \mathbf{w}_2], \mathbf{w}_1, \mathbf{w}_2 \in C$ then $\mathbf{w} = \mathbf{w}_1 = \mathbf{w}_2$. For a convex set C, $\mathbf{w} \in C$ is an extreme point iff $C \setminus \{\mathbf{w}\}$ remains convex.

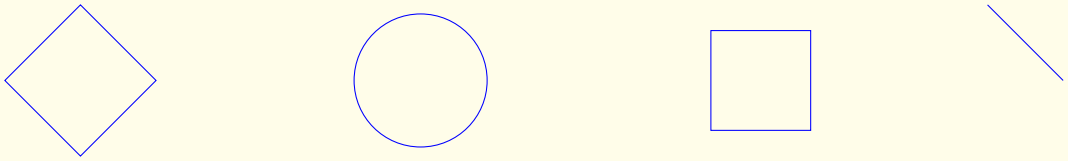
Prove that for any convex function, the maximizer in

$$\operatorname*{argmax}_{\mathbf{w}\in C} f(\mathbf{w})$$

can always be chosen as an extreme point of C. [Hint: recall the definition of a convex function.]

Exercise 6.9: Extreme points of balls

Find out the extreme points of the ℓ_p-norm ball $\mathsf{B}_p := \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\|_p \le 1\}$, for $p \in \{1, 2, \infty\}$, as well as the simplex $\Delta := \{\mathbf{w} \in \mathbb{R}^d : \mathbf{w} \ge \mathbf{0}, \sum_j w_j = 1\}$. The following pictures in $\mathbb{R}^2$ may help you:



[Hint: read the first paragraph in Exercise 6.8.]

Remark 6.10: Step size selection

There are a few ways to choose the step size:

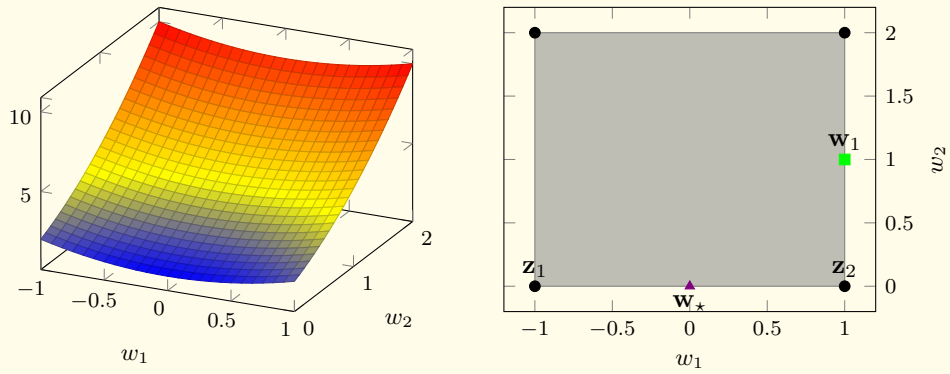
- Open-loop rule: $\eta_t = \frac{2}{t+2}$, or more generally $\eta_t = \Theta(1/t)$.
- Cauchy's rule: $\eta_t \in \operatorname*{argmin}_{0 \le \eta \le 1} f((1-\eta)\mathbf{w}_t + \eta \mathbf{z}_t)$.
- Quadratic rule: $\eta_t = \operatorname*{argmin}_{0 \le \eta \le 1} f(\mathbf{w}_t) + \eta \langle \mathbf{z}_t - \mathbf{w}_t; \nabla f(\mathbf{w}_t) \rangle + \frac{\mathsf{L}^2 \eta_t^2 \|\mathbf{w}_t - \mathbf{z}_t\|^2}{2} = \left[\frac{\langle \mathbf{w}_t - \mathbf{z}_t; \nabla f(\mathbf{w}_t) \rangle}{\mathsf{L}^2 \|\mathbf{w}_t - \mathbf{z}_t\|^2} \right]_0^1$.

Example 6.11: Does it work?

Consider the following simple problem:

$$\min_{\mathbf{w} \in C} f(\mathbf{w}), \quad \text{where} \quad f(\mathbf{w}) = w_1^2 + (w_2 + 1)^2 \quad \text{and} \quad C := \{\mathbf{w} : w_1 \in [-1, 1], w_2 \in [0, 2]\}.$$

The global minimizer is clearly at $\mathbf{w}_\star = (0, 0)$, as indicated in the following plots:



Let us see how the conditional gradient Algorithm 6.7 works on this toy problem:

- We first identify the four extreme points of C as

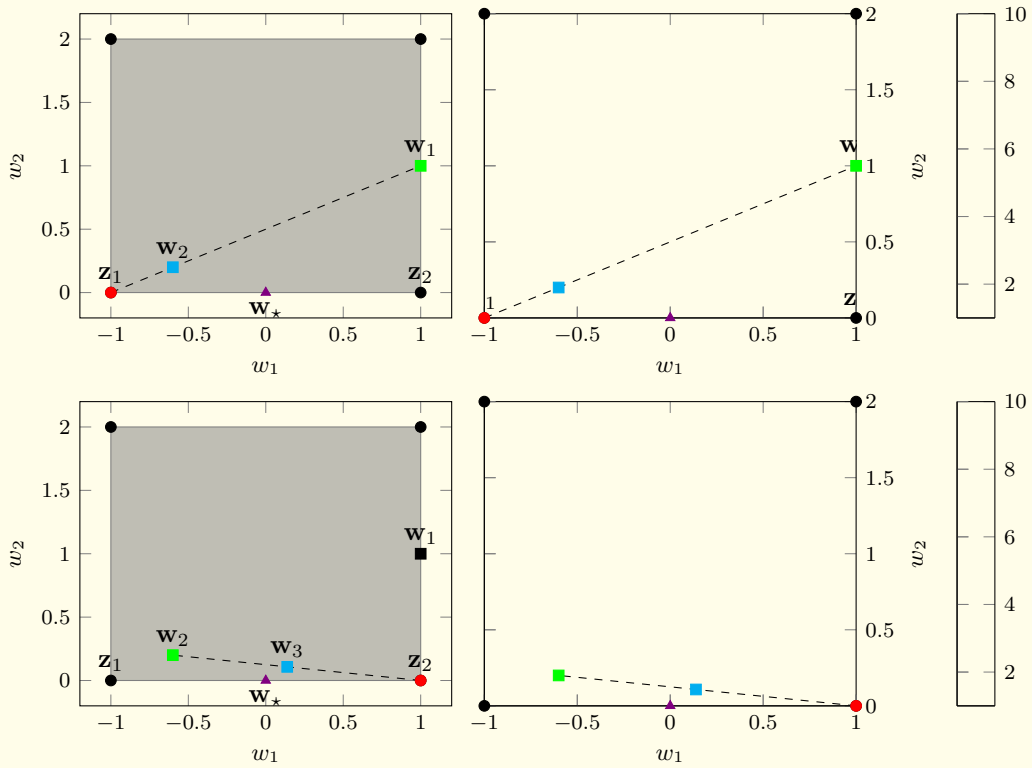
$$\mathbf{z}_1 = (-1, 0), \quad \mathbf{z}_2 = (1, 0), \quad \mathbf{z}_3 = (1, 2), \quad \mathbf{z}_4 = (-1, 2).$$

- Start with say $\mathbf{w}_1 = (1, 1)$, we compute the gradient $\nabla f(\mathbf{w}_1) = (2, 4)$.
- Then, we pick the extreme point $\mathbf{z}$ that maximizes $\langle \mathbf{z}; -\nabla f(\mathbf{w}_1) \rangle$. Clearly, $\mathbf{z}_1$ wins.

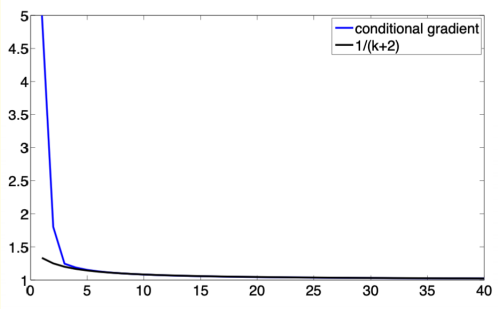
- Next, we find $\eta > 0$ to minimize $f((1-\eta)\mathbf{w}_1 + \eta\mathbf{z}_1)$ by setting its derivative w.r.t. η to 0 :

$$\eta_1 = \eta = \frac{\langle \mathbf{w} + (0,1), \mathbf{w} - \mathbf{z} \rangle}{\|\mathbf{w} - \mathbf{z}\|_2^2} = \frac{4}{5}.$$

- Lastly, we compute $\mathbf{w}_2 = (1-\eta_1)\mathbf{w}_1 + \eta_1\mathbf{z}_1 = (-\frac{3}{5}, \frac{1}{5})$, and the process repeats.



The convergence rate closely follows $\Theta(1/t)$, while projected gradient Algorithm 3.15 converges in 2 iterations for this example!



Remark 6.12: Greedy algorithm for submodular functions

If C is the base polyhedron of a submodular function, then the linear subproblem is exactly Edmonds' greedy algorithm.

History 6.13: Frank and Wolfe's motivation

The original motivation behind Algorithm 6.7 is to solve quadratic programs (see also Wolfe (1959)), using the then-newly-minted simplex algorithm for linear programming (LP) as a subroutine. Indeed, when C is a polytope, the linear subproblem in line 2 is an LP.

We note that George Dantzig's PhD thesis was on solving an open problem that he mistakenly thought to be an assignment (since he was late for class)...

Wolfe, P. (1959). "The Simplex Method for Quadratic Programming". *Econometrica*, vol. 27, no. 3, pp. 382–398.

Example 6.14: Sparsity! (e.g. Shalev-Shwartz et al. 2010; Clarkson 2010; Hazan 2008)

Let $C := \{\mathbf{w} : \|\mathbf{w}\|_1 \le \lambda\}$ be the scaled ℓ_1 norm ball. Then, the polar operator in line 2 of Algorithm 6.7, with the help of Exercise 6.9, reduces to

$$\mathbf{z}_t = \operatorname*{argmax}_{\|\mathbf{z}\|_1 \le \lambda} \langle \mathbf{z}; -\nabla f(\mathbf{w}_t)\rangle \ni -\lambda \mathbf{e}_i, \quad \text{where} \quad \langle \mathbf{e}_i; \nabla_i f(\mathbf{w}_t)\rangle = \max_j |\nabla_j f(\mathbf{w}_t)|.$$

In particular, we may choose $\mathbf{e}_i$ to be the i-th standard basis (i.e. 1 at the i-th entry and 0 elsewhere). It follows that after t steps, the iterate $\mathbf{w}_t$ has (added) at most t nonzeros! In comparison, after even a single iteration, projected gradient can result in a fully dense iterate! On the downside, the resulting coordinate-wise update is a bit wasteful though: we compute the entire gradient ∇f only to find its minimum index and throw out everything else...

The picture changes dramatically once we move to the matrix setting; recall Example 6.3. With $C = \{W : \|W\|_{\mathrm{tr}} \le \lambda\}$ being the trace norm ball, the polar operator reduces to computing the spectral norm:

$$Z_t = \operatorname*{argmax}_{\|Z\|_{\mathrm{tr}} \le \lambda} \langle Z; -\nabla f(W_t)\rangle = -\lambda \mathbf{u}\mathbf{v}^\top, \quad \text{where} \quad \mathbf{u}^\top \nabla f(W_t)\mathbf{v} = \|\nabla f(W_t)\|_{\mathrm{sp}}.$$

Thus, after t steps, the iterate W_t has (added) rank at most t. Moreover, computing the spectral norm, i.e. the largest singular value, costs quadratic time, which is one magnitude cheaper than the projection operator (which costs cubic time due to computing all singular values).

Shalev-Shwartz, S., N. Srebro, and T. Zhang (2010). "Trading Accuracy for Sparsity in Optimization Problems with Sparsity Constraints". *SIAM Journal on Optimization*, vol. 20, no. 6, pp. 2807–2832.

Clarkson, K. L. (2010). "Coresets, Sparse Greedy Approximation, and the Frank-Wolfe Algorithm". *ACM Transactions on Algorithms*, vol. 6, no. 4, pp. 1–30.

Hazan, E. (2008). "Sparse Approximate Solutions to Semidefinite Programs". In: *Latin American Conference on Theoretical Informatics*, pp. 306–316.

Remark 6.15: Algorithmic equivalence between projection and polar

Despite the drastic difference in complexity, the projection operator and the polar operator are in some sense algorithmically equivalent: Given a polar operator, we can simply apply the conditional gradient algorithm to numerically compute the projection. The converse is also true using a simple bisection search, see e.g. Zhang et al. (2013, Proposition 2).

Zhang, X., Y. Yu, and D. Schuurmans (2013). "Polar Operators for Structured Sparse Estimation". In: *Advances in Neural Information Processing Systems 27 (NIPS)*.

Theorem 6.16: Convergence in terms of function values

Suppose f is convex and $\mathsf{L} = \mathsf{L}^{[1]}$-smooth, and C is compact convex with bounded diameter ρ. Then, the

iterates generated by Algorithm 6.7 satisfy: for all $\mathbf{w} \in C$,

$$f(\mathbf{w}_{t+1}) \le f(\mathbf{w}) + \pi_t(1-\eta_0)(f(\mathbf{w}_0) - f(\mathbf{w})) + \frac{\mathsf{L}\rho^2}{2}\sum_{s=0}^{t}\frac{\pi_t}{\pi_s}\eta_s^2, \tag{6.3}$$

where $\pi_t := \prod_{s=1}^{t}(1-\eta_s)$ *with* $\pi_0 := 1$. *In particular, setting* $\eta_t = \frac{2}{t+2}$, *we have* $\eta_0 = 1$, $\pi_t = \frac{2}{(t+1)(t+2)}$ *and*

$$f(\mathbf{w}_t) - f(\mathbf{w}) \le \frac{2\mathsf{L}\rho^2}{t+3},$$

where the initializer $\mathbf{w}_0$, *surprisingly, does not play any role.*

Proof: Indeed, for any $\mathbf{w} \in C$:

$$\begin{aligned}
f(\mathbf{w}_{t+1}) - f(\mathbf{w}) &= f((1-\eta_t)\mathbf{w}_t + \eta_t\mathbf{z}_t) - f(\mathbf{w}) \\
(\mathsf{L}^{[1]}\text{-smoothness}) \quad &\le f(\mathbf{w}_t) - f(\mathbf{w}) + \eta_t\langle \mathbf{z}_t - \mathbf{w}_t; \nabla f(\mathbf{w}_t)\rangle + \frac{\eta_t^2}{2}\mathsf{L}\underbrace{\|\mathbf{w}_t - \mathbf{z}_t\|^2}_{\le \rho^2} \\
(\text{optimality of } \mathbf{z}_t) \quad &\le f(\mathbf{w}_t) - f(\mathbf{w}) + \eta_t\langle \mathbf{w} - \mathbf{w}_t; \nabla f(\mathbf{w}_t)\rangle + \frac{\eta_t^2}{2}\mathsf{L}\rho^2 \\
(\text{convexity of } f) \quad &\le (1-\eta_t)(f(\mathbf{w}_t) - f(\mathbf{w})) + \frac{\eta_t^2}{2}\mathsf{L}\rho^2.
\end{aligned}$$

Telescoping and collecting the terms we arrive at the claim (6.3). ■

In fact, with the same step size choice, we can prove the same convergence rate for the duality gap:

$$f(\mathbf{w}_t) - f(\mathbf{w}) \le \langle \mathbf{w}_t - \mathbf{w}; \nabla f(\mathbf{w}_t)\rangle \le \langle \mathbf{w}_t - \mathbf{z}_t; \nabla f(\mathbf{w}_t)\rangle \le \frac{2\mathsf{L}\rho^2}{t+3}, \tag{6.4}$$

where the highlighted term is computable hence can be monitored as a convergence criteria.

Inspecting the proof we see that all we need about f, aside from convexity and l.s.c., is that there exists some constant K so that for all $\mathbf{w}, \mathbf{z} \in C$ and $\eta \in [0,1]$,

$$f((1-\eta)\mathbf{w} + \eta\mathbf{z}) \le f(\mathbf{w}) + \eta\langle \mathbf{z} - \mathbf{w}; \nabla f(\mathbf{w})\rangle + K\eta^2,$$

which continues to make sense even in a TVS, as already being pointed out in e.g. Auslender (1968).

Auslender, A. (1968). "Algorithme de recherche des points stationnaires d'une fonctionnelle dans un espace vectoriel topologique application a un probleme de controle a evolution non lineaire". *Comptes rendus mathématiques de l'Académie des Sciences, Paris*, vol. 266, pp. 226–229.

Example 6.17: Lower bound in high dimension

We mention the simple *lower bound* from Jaggi (2013). Consider the trivial minimization problem (that vacuously projects a point $\frac{1}{d}\mathbf{1}$ to the simplex where it resides):

$$\min_{\mathbf{w}\in\Delta} \ \tfrac{1}{2}\|\mathbf{w} - \tfrac{1}{d}\mathbf{1}\|_2^2.$$

Let us apply Algorithm 6.7 with initialization $\mathbf{w}_0 = \mathbf{e}_1$. Due to the sparse property (see Example 6.14), after t iterations, $\mathbf{w}_t$ has at most $t+1$ nonzeros. However, using symmetry we know that

$$O(\tfrac{1}{t}) = \tfrac{1}{2}(\tfrac{1}{t} - \tfrac{1}{d}) = \tfrac{1}{2}[t(\tfrac{1}{t} - \tfrac{1}{d})^2 + (d-t)(\tfrac{1}{d})^2] = \min_{\mathbf{w}\in\Delta, \|\mathbf{w}\|_0 = t} \tfrac{1}{2}\|\mathbf{w} - \tfrac{1}{d}\mathbf{1}\|_2^2 = \tfrac{1}{2}\|\mathbf{w}\|_2^2 - \tfrac{1}{d} + \tfrac{1}{2d}.$$

In other words, the conditional gradient algorithm converges exactly at the rate $\Theta(\frac{1}{t})$ for this simple problem (when the dimension d is large).

Jaggi, M. (2013). "Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization". In: *Proceedings of the 30th International Conference on Machine Learning*, pp. 427–435.

Example 6.18: Exact $\Theta(1/t)$ rate in low dimension (Canon and Cullum 1968)

Consider the simple 2-d minimization problem:

$$\min_{a,b} \ a^2 + (1+b)^2, \quad \text{s.t.} \ \ |a| \le 1, \ 1 \ge b \ge 0.$$

The following observations are easy to verify:

- The minimum value $f_\star = 1$ is attained at a unique minimizer $\mathbf{w}_\star = (0, 0)$, which lies on the boundary of the constraint.
- Line 2 yields $\mathbf{z}_t = (-\operatorname{sign}(a_t), 0)$, since the gradient is $(2a_t, 2(1+b_t))$.
- Line 4 chooses η optimally: $\eta_t = \operatorname*{argmin}_{\eta \in [0,1]} [(1-\eta)|a_t| - \eta]^2 + [1 + (1-\eta)b_t]^2 = \frac{b_t + |a_t| + a_t^2 + b_t^2}{(|a_t|+1)^2 + b_t^2} \wedge 1$.
- For t large, $\eta_t < 1$ (since $a_t, b_t \to 0$ according to Theorem 6.16) and

$$f_{t+1} = f_t - \frac{(a_t^2 + |a_t| + b_t^2 + b_t)^2}{(|a_t|+1)^2 + b_t^2}.$$

- It can be proved that $f_t = \Omega(1/t)$, while from Theorem 6.16 we already know $f_t = O(1/t)$ hence in fact $f_t = \Theta(1/t)$.

We have thus created a simple low dimensional example where Algorithm 6.7 converges exactly at $\Theta(1/t)$. Put differently, the rate proved in Theorem 6.16 cannot be improved without further assumption or modification! We note that a more substantial lower bound that also holds for other algorithms is given in Guzmán and Nemirovski (2015).

Canon, M. D. and C. D. Cullum (1968). "Tight Upper Bound on the Rate of Convergence of Frank-Wolfe Algorithm". *SIAM Journal on Control*, vol. 6, no. 4, pp. 509–516.
Guzmán, C. and A. Nemirovski (2015). "On lower complexity bounds for large-scale smooth convex optimization". *Journal of Complexity*, vol. 31, pp. 1–14.

Remark 6.19: Multiplicative approximate guarantee

Recall from Example 6.14 that when we lift from vectors to matrices, conditional gradient saved us an order of magnitude per-step complexity. Things get even more weary if we lift the domain further to tensors (e.g. high dimensional matrices): the tensor spectral or trace norm is known to be NP-hard to evaluate. Nevertheless, there is a straightforward multiplicative algorithm for approximating the tensor spectral norm, and Cheng et al. (2016) showed that the conditional gradient algorithm pleasantly inherits the same multiplicative factor (which unfortunately is dimension dependent).

Cheng, H., Y. Yu, X. Zhang, E. Xing, and D. Schuurmans (2016). "Scalable and Sound Low-Rank Tensor Learning". In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Algorithm 6.20: Extension to the composite setting

For the composite minimization problem

$$\min_{\mathbf{w}} \ \ell(\mathbf{w}) + r(\mathbf{w}),$$

Bonesky et al. (2007) and Bredies et al. (2009) proposed the generalized conditional gradient algorithm,

which amounts to changing Line 2 in Algorithm 6.7 naturally to

$$\mathbf{z}_t = \operatorname*{argmin}_{\mathbf{z}} \langle \mathbf{z}; \nabla \ell(\mathbf{w}_t) \rangle + r(\mathbf{z}).$$

When r is homogeneous (e.g. a norm), the above subroutine may not admit any minimizer (diverging possibly to $-\infty$), which was addressed by Yu et al. (2017) through a simple line search.

Bonesky, T., K. Bredies, D. A. Lorenz, and P. Maass (2007). "A Generalized Conditional Gradient Method for Nonlinear Operator Equations with Sparsity Constraints". *Inverse Problems*, vol. 23, no. 5, pp. 2041–2058.
Bredies, K., D. A. Lorenz, and P. Maass (2009). "A Generalized Conditional Gradient Method and its Connection to an Iterative Shrinkage Method". *Computational Optimization and Applications*, vol. 42, pp. 173–193.
Yu, Y., X. Zhang, and D. Schuurmans (2017). "Generalized Conditional Gradient for Structured Sparse Estimation". *Journal of Machine Learning Research*, vol. 18, pp. 1–46.

Definition 6.21: Weakly and strongly convex functions

We call a function $f : \mathbb{R}^d \to \mathbb{R}$ σ-weakly convex (w.r.t. some norm $\|\cdot\|$) if there exists some $\sigma \in \mathbb{R}$ such that for all $\mathbf{w}, \mathbf{z} \in \mathbb{R}^d$ and $\lambda \in [0,1]$:

$$f(\lambda \mathbf{z} + (1-\lambda)\mathbf{w}) + \sigma \cdot \frac{\lambda(1-\lambda)}{2}\|\mathbf{w}-\mathbf{z}\|^2 \le \lambda f(\mathbf{z}) + (1-\lambda) f(\mathbf{w}).$$

The largest such $\sigma = \sigma(f)$ is called the modulus of convexity of f.

Some authors call f σ-strongly convex when $\sigma > 0$.

Proposition 6.22: Characterization of weak and strong convexity

A smooth function $f : \mathbb{R}^d \to \mathbb{R}$ is σ-weakly convex iff

(I). $\forall \mathbf{w}, \forall \mathbf{z},\ f(\mathbf{z}) \ge f(\mathbf{w}) + \langle \mathbf{z}-\mathbf{w}; \nabla f(\mathbf{w}) \rangle + \frac{\sigma}{2}\|\mathbf{w}-\mathbf{z}\|^2$.

(II). $\forall \mathbf{w}, \forall \mathbf{z},\ \langle \mathbf{z}-\mathbf{w}; \nabla f(\mathbf{z}) - \nabla f(\mathbf{w}) \rangle \ge \sigma \|\mathbf{w}-\mathbf{z}\|^2$.

(III). $\forall \mathbf{w}, \forall \mathbf{z},\ \langle \mathbf{z}; \nabla^2 f(\mathbf{w})\mathbf{z} \rangle \ge \sigma \|\mathbf{z}\|^2$.

For strongly convex functions (i.e. $\sigma(f) > 0$) we further have the following equivalence:

(IV). $\forall \mathbf{w}, \forall \mathbf{z},\ f(\mathbf{z}) \le f(\mathbf{w}) + \langle \mathbf{z}-\mathbf{w}; \nabla f(\mathbf{w}) \rangle + \frac{1}{2\sigma}\|\nabla f(\mathbf{w}) - \nabla f(\mathbf{z})\|_\circ^2$.

(V). $\forall \mathbf{w}, \forall \mathbf{z},\ \langle \mathbf{z}-\mathbf{w}; \nabla f(\mathbf{z}) - \nabla f(\mathbf{w}) \rangle \le \frac{1}{\sigma}\|\nabla f(\mathbf{w}) - \nabla f(\mathbf{z})\|_\circ^2$.

(VI). $\forall \mathbf{w}, \forall \mathbf{z},\ \|\nabla f(\mathbf{z}) - \nabla f(\mathbf{w})\|_\circ \ge \sigma \|\mathbf{w}-\mathbf{z}\|$.

Proof: The proofs for (I) $\iff$ (II) $\iff$ (III) are completely similar to Proposition 2.12, while the implications (II) $\implies$ (VI) $\implies$ (V) and (IV) $\implies$ (V) are clear.

Let us first prove (V) $\implies$ (III). Indeed, dividing $\|\mathbf{z}-\mathbf{w}\|^2$ on both sides of (V) and letting $\frac{\mathbf{z}-\mathbf{w}}{\|\mathbf{z}-\mathbf{w}\|} \to \mathbf{u}$ we have

$$\langle \mathbf{u}; \nabla^2 f(\mathbf{w})\mathbf{u} \rangle \le \frac{1}{\sigma}\|\nabla^2 f(\mathbf{w})\mathbf{u}\|_\circ^2.$$

Since f is strongly convex, we know from (III) that $\nabla^2 f$ is non-degenerate. Conjugating on both sides above we obtain:

$$\langle \mathbf{u}^*; [\nabla^2 f(\mathbf{w})]^{-1}\mathbf{u}^* \rangle \ge \sigma \|[\nabla^2 f(\mathbf{w})]^{-1}\mathbf{u}^*\|^2.$$

which, upon change of variable, is exactly (III).

Lastly, we prove (I) $\implies$ (IV) when $\sigma \ge 0$. Fixing $\mathbf{w}$, it is clear that

$$\mathsf{D}_f(\mathbf{v}; \mathbf{w}) := f(\mathbf{v}) - f(\mathbf{w}) - \langle \mathbf{v}-\mathbf{w}; \nabla f(\mathbf{w}) \rangle$$

is also σ-weakly convex as a function of $\mathbf{v}$, i.e.

$$\mathsf{D}_f(\mathbf{v};\mathbf{w}) \geq \mathsf{D}_f(\mathbf{z};\mathbf{w}) + \langle \mathbf{v}-\mathbf{z}; \nabla f(\mathbf{z}) - \nabla f(\mathbf{w})\rangle + \tfrac{\sigma}{2}\|\mathbf{v}-\mathbf{z}\|^2,$$

which can also be directly verified from (I) by straightforward simplification. Minimizing w.r.t. $\mathbf{v}$ on both sides and using the fact that $\mathsf{D}_f \geq 0$ with 0 achievable (at say $\mathbf{v}=\mathbf{w}$):

$$\begin{aligned} 0 &\geq \min_{\mathbf{v}} \mathsf{D}_f(\mathbf{z};\mathbf{w}) + \langle \mathbf{v}-\mathbf{z}; \nabla f(\mathbf{z}) - \nabla f(\mathbf{w})\rangle + \tfrac{\sigma}{2}\|\mathbf{v}-\mathbf{z}\|^2 \\ &= \min_{t} \mathsf{D}_f(\mathbf{z};\mathbf{w}) + \|\nabla f(\mathbf{z}) - \nabla f(\mathbf{w})\|_\circ t + \tfrac{\sigma}{2}t^2 \\ &= \mathsf{D}_f(\mathbf{z};\mathbf{w}) - \tfrac{1}{2\sigma}\|\nabla f(\mathbf{z}) - \nabla f(\mathbf{w})\|_\circ^2, \end{aligned}$$

which is exactly (IV). ■

In particular, when the underlying norm is Euclidean, condition (III) means that the smallest eigenvalue of the Hessian is bounded from below by σ, or equivalently iff $f - \frac{\sigma}{2}\|\cdot\|_2^2$ is convex. However, we emphasize that this neat observation only holds for the Euclidean norm! Sun and Yu (2019) and the references therein documented many nice properties of weakly convex functions.

We point out that (IV) and (V) trivially hold for any constant function (with any σ), which is why we have to restrict to strongly convex functions in establishing the equivalence (V) $\implies$ (III). On the other hand, (VI) automatically implies the function is strongly convex.

Rockafellar (1976, Prop 6) established the equivalence to (I) and (II), even when f may be nonsmooth.

Sun, S. and Y. Yu (2019). "Least Squares Estimation of Weakly Convex Functions". In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Rockafellar, R. T. (1976). "Monotone Operators and the Proximal Point Algorithm". *SIAM Journal on Control and Optimization*, vol. 14, no. 5, pp. 877–898.

Exercise 6.23: Spectrum equality

Prove the following equality for any positive definite matrix A:

$$\min_{\|\mathbf{u}\|=1} \langle \mathbf{u}; A\mathbf{u}\rangle = \min_{\|\mathbf{u}\|=1} \|A\mathbf{u}\|_\circ.$$

(The similar equality relating the maximum was already established in the proof of Theorem 2.13). This result is easy for the Euclidean norm. However, we are asking for a proof that works for any norm.

[Hint: consider the strongly convex function $\frac{1}{2}\langle \mathbf{u}; A\mathbf{u}\rangle$. Can you find a direct proof?]

Example 6.24: σ-strong convexity and $\mathsf{L}^{[1]}$-smoothness

Consider the univariate function

$$f(w) = \sin(w) + w^2.$$

It is easy to verify that

$$f'(w) = \cos(w) + 2w, \qquad f''(w) = -\sin(w) + 2 \in [1,3].$$

Thus, we have

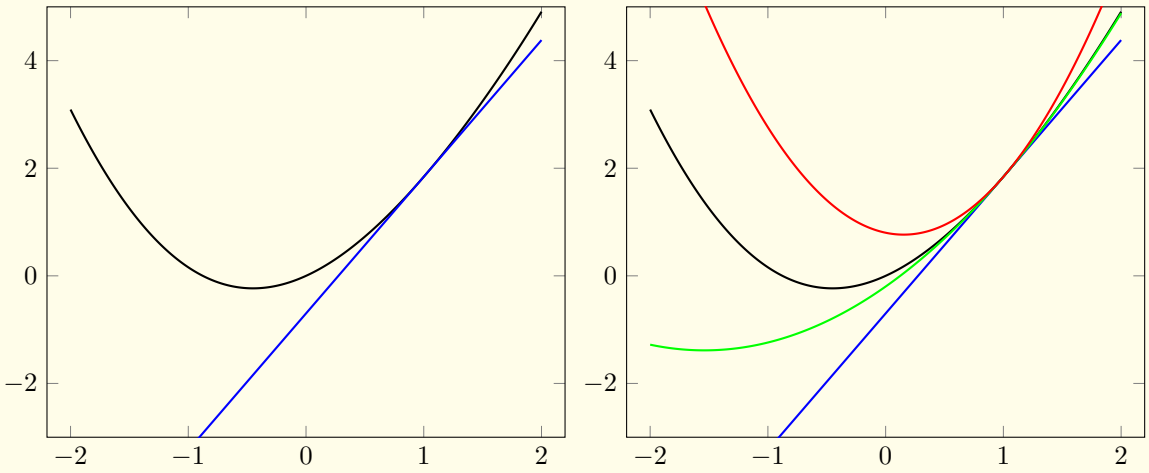
$$\sigma = \left[\inf_w \ f''(w)\right] = 1, \qquad \mathsf{L} = \mathsf{L}^{[1]} = \left[\sup_w \ f''(w)\right] = 3.$$

Fix any z, for definiteness say $z=1$. Since f is convex, we have the lower bound (see Theorem 0.29):

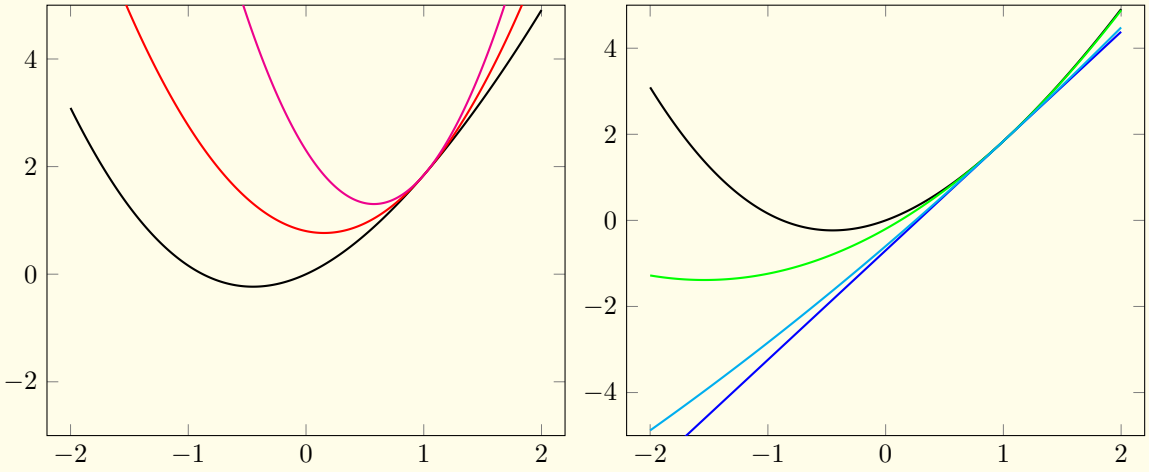
$$f(w) \geq \underbrace{f(z) + f'(z)(w-z)}_{:=\ell(w)} = \sin(1) + 1 + [\cos(1)+2](w-1).$$

In fact, we can strengthen the bound using σ-convexity and L-smoothness:

$$\ell(w) + \tfrac{\sigma}{2}(w-z)^2 \le f(w) \le \ell(w) + \tfrac{\mathsf{L}}{2}(w-z)^2.$$



The black curve above depicts our function f while the blue straight line is the linear lower bound $\ell(w)$. On the right panel, the red curve, with $\mathsf{L} = 3$, gives the tightest upper quadratic bound $\ell(w) + \frac{\mathsf{L}}{2}(w-z)^2$ while the green curve, with $\sigma = 1$, gives the tightest lower quadratic bound $\ell(w) + \frac{\sigma}{2}(w-z)^2$. Note that all curves coincide at z, the point we choose to derive the bounds.



Of course, any $\mathsf{L} \ge 3$ is also a valid $\mathsf{L}^{[1]}$-smoothness parameter (just not the tightest). On the left panel above, with a bigger L, e.g. $\mathsf{L} = 6$, the quadratic upper bound $\ell(w) + \frac{\mathsf{L}}{2}(w-z)^2$ becomes steeper (the pink curve) when compared to $\mathsf{L} = 3$ (the red curve). It is apparent that with a bigger L the quadratic bound is looser and steeper, hence the range where it approximates f well becomes smaller. This explains why in gradient algorithms we usually have $\eta \propto \frac{1}{\mathsf{L}}$, i.e., the bigger L is, the smaller the step size needs to be in order to avoid moving too fast (hence stepping outside of the reasonable range of approximation). On the right panel, we show the two quadratic lower bounds ($\sigma = 0.2$ vs. $\sigma = 1$). It is clear that any $\sigma \le 1$ is also a valide convexity parameter (just not the tightest). As $\sigma \to 0$, the quadratic lower bound (in cyan) becomes looser and looser, and eventually approaches the linear lower bound (in blue, $\sigma = 0$).

Alert 6.25: Duality between smoothness and strong convexity

The equivalent conditions (I) and (IV) in Proposition 6.22 actually reveal something fundamental: A convex function f is σ-strongly convex iff its Fenchel conjugate function f^* (see Definition 0.28) is $\frac{1}{\sigma}$-smooth!

To see this equivalence, let us recall the Fenchel-Young equality:

$$f(\mathbf{w}) + f^*(\nabla f(\mathbf{w})) = \langle \mathbf{w}; \nabla f(\mathbf{w}) \rangle, \text{ and } \mathbf{z} = \nabla f^*(\nabla f(\mathbf{z})).$$

Thus, condition (IV) can be rewritten as

$$-f^*(\nabla f(\mathbf{z})) \le -f^*(\nabla f(\mathbf{w})) + \langle \mathbf{z}; \nabla f(\mathbf{w}) - \nabla f(\mathbf{z}) \rangle + \tfrac{1/\sigma}{2} \|\nabla f(\mathbf{w}) - \nabla f(\mathbf{z})\|_\circ^2,$$

which is exactly the $\frac{1}{\sigma}$-smoothless of f^*!

It is instructive to simplify the conditions in Proposition 6.22, Theorem 2.13, Proposition 2.12 and Exercise 6.26 (below) for the quadratic function $\mathsf{q}_A(\mathbf{u}) = \frac{1}{2}\langle \mathbf{u}; A\mathbf{u} \rangle$. Inspecting the proofs we realize that the duality between smoothness and strong convexity is largely a property of positive definite matrices (that induce quadratic functions $\langle \mathbf{u}; A\mathbf{u} \rangle$), along with the simple conjugation result: $[\frac{1}{2}\|\cdot\|^2]^* = \frac{1}{2}\|\cdot\|_\circ^2$.

Exercise 6.26: Duality on smoothness

A convex function f is $\mathsf{L} = \mathsf{L}^{[1]}$-smooth iff for all $\mathbf{w}$ and $\mathbf{z}$:

- $f(\mathbf{z}) \ge f(\mathbf{w}) + \langle \mathbf{z} - \mathbf{w}; \nabla f(\mathbf{w}) \rangle + \frac{1}{2\mathsf{L}} \|\nabla f(\mathbf{z}) - \nabla f(\mathbf{w})\|_\circ^2$.
- $\langle \mathbf{z} - \mathbf{w}; \nabla f(\mathbf{z}) - \nabla f(\mathbf{w}) \rangle \ge \frac{1}{\mathsf{L}} \|\nabla f(\mathbf{w}) - \nabla f(\mathbf{z})\|_\circ^2$.

Definition 6.27: Matching loss

Let us record the simple equality:

$$\mathsf{D}_f(\mathbf{w}; \mathbf{z}) = \mathsf{D}_{f^*}(\nabla f(\mathbf{z}); \nabla f(\mathbf{w})), \text{ note the swapping of arguments.}$$

L-smoothness and σ-weak convexity simply correspond to the left and right side of the sandwiching:

$$\tfrac{\mathsf{L}}{2}\|\mathbf{w} - \mathbf{z}\|^2 \ge \mathsf{D}_f(\mathbf{w}; \mathbf{z}) \ge \tfrac{\sigma}{2}\|\mathbf{w} - \mathbf{z}\|^2,$$

which, using duality and the above equality, is equivalent to

$$\tfrac{1}{2\sigma}\|\nabla f(\mathbf{z}) - \nabla f(\mathbf{w})\|^2 \ge \mathsf{D}_f(\mathbf{w}; \mathbf{z}) = \mathsf{D}_{f^*}(\nabla f(\mathbf{z}); \nabla f(\mathbf{w})) \ge \tfrac{1}{2\mathsf{L}}\|\nabla f(\mathbf{z}) - \nabla f(\mathbf{w})\|^2.$$

Theorem 6.28: Linear convergence (Guélat and Marcotte 1986; Beck and Teboulle 2004)

Suppose f is σ-strongly convex and $\mathsf{L} = \mathsf{L}^{[1]}$-smooth, C is closed convex with diameter ρ, and the unique minimizer $\mathbf{w}_\star$ is in the interior, i.e. $\mathsf{B}_{\mathbf{w}_\star}(\mathsf{r}) \subseteq C$. Then, the iterates generated by Algorithm 6.7, with either Cauchy's or quadratic step size, satisfy:

$$\forall t \ge T := 4\gamma - 3, \quad f(\mathbf{w}_t) - f(\mathbf{w}_\star) \le (1 - \gamma)^{t-T} \left(f(\mathbf{w}_T) - f(\mathbf{w}_\star)\right),$$

i.e., the algorithm converges linearly, where the "condition number" $\gamma := \frac{\sigma \mathsf{r}^2}{\mathsf{L}\rho^2}$.

Proof: We bound the progress of the algorithm similarly as in Theorem 6.16:

$$\begin{aligned} f(\mathbf{w}_{t+1}) - f(\mathbf{w}_\star) &= f((1 - \eta_t)\mathbf{w}_t + \eta_t \mathbf{z}_t) - f(\mathbf{w}_\star) \\ (L\text{-smoothness}) \ &\le f(\mathbf{w}_t) - f(\mathbf{w}_\star) + \eta_t \langle \mathbf{z}_t - \mathbf{w}_t; \nabla f(\mathbf{w}_t) \rangle + \frac{\eta_t^2}{2}\mathsf{L}\|\mathbf{w}_t - \mathbf{z}_t\|^2 \end{aligned}$$

$$\begin{aligned}
\text{(optimality of } \mathbf{z}_t \text{ vs. } \mathsf{B}_{\mathbf{w}_\star}(\mathsf{r})) \quad &\le f(\mathbf{w}_t) - f(\mathbf{w}_\star) + \eta_t \langle \mathbf{w}_\star - \mathbf{w}_t; \nabla f(\mathbf{w}_t)\rangle - \eta_t \mathsf{r}\|\nabla f(\mathbf{w}_t)\|_\circ + \frac{\eta_t^2 \mathsf{L}}{2}\rho^2 \\
\text{(convexity of } f) \quad &\le f(\mathbf{w}_t) - f(\mathbf{w}_\star) - \eta_t \mathsf{r}\|\nabla f(\mathbf{w}_t)\|_\circ + \frac{\eta_t^2 \mathsf{L}}{2}\rho^2 \\
&\le f(\mathbf{w}_t) - f(\mathbf{w}_\star) - \eta_t \mathsf{r}\sqrt{2\sigma(f(\mathbf{w}_t) - f(\mathbf{w}_\star))} + \frac{\eta_t^2 \mathsf{L}}{2}\rho^2,
\end{aligned}$$

where the last inequality follows from condition (IV) in Proposition 6.22 (recall that $\nabla f(\mathbf{w}_\star) = \mathbf{0}$ since $\mathbf{w}_\star$ is an interior point). Choosing $\eta_t \in [0, 1]$ to minimize the last term:

$$\eta_t = \frac{\mathsf{r}\sqrt{2\sigma(f(\mathbf{w}_t) - f_\star)}}{\mathsf{L}\rho^2} \wedge 1,$$

which yields the estimate

$$f(\mathbf{w}_{t+1}) - f_\star \le \left(1 - \tfrac{\sigma \mathsf{r}^2}{\mathsf{L}\rho^2}\right)(f(\mathbf{w}_t) - f_\star),$$

as long as $\eta_t \le 1$, i.e., $t \ge T$ using the sublinear rate (6.4) in Theorem 6.16. ■

The linear convergence rate here was first proved by Guélat and Marcotte (1986) and then rediscovered in Beck and Teboulle (2004). Garber and Hazan (2015) also proved a faster rate when the constraint is "strongly convex," while Ahipasaoglu et al. (2008) and Beck and Shtern (2017) contain further refinements.

Guélat, J. and P. Marcotte (1986). "Some comments on Wolfe's 'away step'". *Mathematical Programming*, vol. 35, pp. 110–119.
Beck, A. and M. Teboulle (2004). "A Conditional Gradient Method with Linear Rate of Convergence for Solving Convex Linear Systems". *Mathematical Methods of Operations Research*, vol. 59, pp. 235–247.
Garber, D. and E. Hazan (2015). "Faster Rates for the Frank-Wolfe Method over Strongly-Convex Sets". In: *Proceedings of the 32nd International Conference on Machine Learning*, pp. 541–549.
Ahipasaoglu, S. D., P. Sun, and M. J. Todd (2008). "Linear convergence of a modified Frank–Wolfe algorithm for computing minimum-volume enclosing ellipsoids". *Optimization Methods and Software*, vol. 23, no. 1, pp. 5–19.
Beck, A. and S. Shtern (2017). "Linearly convergent away-step conditional gradient for non-strongly convex functions". *Mathematical Programming*, vol. 164, pp. 1–27.

Algorithm 6.29: Totally corrective (Meyer 1974; Holloway 1974)

Inspecting Algorithm 6.7 we realize that

$$\mathbf{w}_{t+1} \in \operatorname{conv}\{\mathbf{w}_0, \mathbf{z}_1, \dots, \mathbf{z}_t\},$$

where the extreme points $\mathbf{z}_k$ are repeatedly identified and averaged. One immediate, natural idea is to replace the next iterate $\mathbf{w}_{t+1}$ as the best approximation in the entire convex hull:

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w} \in \operatorname{conv}\{\mathbf{w}_0, \mathbf{z}_1, \dots, \mathbf{z}_t\}} f(\mathbf{w}). \tag{6.5}$$

Such a scheme is termed totally corrective (where we correct all "weak" classifiers), and was due to Meyer (1974) and Holloway (1974). Needless to say, the totally corrective variant can potentially converge much faster (see Example 6.17 though). However, the price to pay is that solving (6.5) is itself non-trivial, even when f is quadratic. In fact, the latter problem has been attacked using the original conditional gradient Algorithm 6.7 in Mitchell et al. (1974) and Wolfe (1976)!

Meyer, G. (1974). "Accelerated Frank–Wolfe Algorithms". *SIAM Journal on Control*, vol. 12, no. 4, pp. 655–655.
Holloway, C. A. (1974). "An extension of the Frank and Wolfe method of feasible directions". *Mathematical Programming*, vol. 6, pp. 14–27.
Mitchell, B. F., V. F. Dem'yanov, and V. N. Malozemov (1974). "Finding the Point of a Polyhedron Closest to the Origin". *SIAM Journal on Control*, vol. 12, no. 1, pp. 19–26.
Wolfe, P. (1976). "Finding the nearest point in A polytope". Vol. 11, pp. 128–149.