Duality between subgradient and conditional gradient methods

Francis Bach

Duality between subgradient and conditional gradient methods

Francis Bach
INRIA - Sierra project-team
Département d'Informatique de l'Ecole Normale Supérieure
Paris, France
francis.bach@ens.fr

October 18, 2013

Abstract

Given a convex optimization problem and its dual, there are many possible first-order algorithms. In this paper, we show the equivalence between mirror descent algorithms and algorithms generalizing the conditional gradient method. This is done through convex duality, and implies notably that for certain problems, such as for supervised machine learning problems with non-smooth losses or problems regularized by non-smooth regularizers, the primal subgradient method and the dual conditional gradient method are formally equivalent. The dual interpretation leads to a form of line search for mirror descent, as well as guarantees of convergence for primal-dual certificates.

1 Introduction

Many problems in machine learning, statistics and signal processing may be cast as convex optimization problems. In large-scale situations, simple gradient-based algorithms with potentially many cheap iterations are often preferred over methods, such as Newton's method or interior-point methods, that rely on fewer but more expensive iterations. The choice of a first-order method depends on the structure of the problem, in particular (a) the smoothness and/or strong convexity of the objective function, and (b) the computational efficiency of certain operations related to the non-smooth parts of the objective function, when it is decomposable in a smooth and a non-smooth part.

In this paper, we consider two classical algorithms, namely (a) subgradient descent and its mirror descent extension [29, 24, 4], and (b) conditional gradient algorithms, sometimes referred to as Frank-Wolfe algorithms [16, 13, 15, 14, 19].

Subgradient algorithms are adapted to non-smooth unstructured situations, and after t steps have a convergence rate of $O(1/\sqrt{t})$ in terms of objective values. This convergence rate improves to $O(1/t)$ when the objective function is strongly convex [22]. Conditional-gradient algorithms are tailored to the optimization of smooth functions on a compact convex set, for which minimizing linear functions is easy (but where orthogonal projections would be hard, so that proximal methods [26, 5] cannot be used efficiently). They also have

a convergence rate of $O(1/t)$ [15]. The main results of this paper are (a) to show that for common situations in practice, these two sets of methods are in fact equivalent by convex duality, (b) to recover a previously proposed extension of the conditional gradient method which is more generally applicable [10], and (c) provide explicit convergence rates for primal and dual iterates. We also review in Appendix A the non-strongly convex case and show that both primal and dual suboptimalities then converge at rate $O(1/\sqrt{t})$.

More precisely, we consider a convex function f defined on $\mathbb{R}^n$, a convex function h defined on $\mathbb{R}^p$, both potentially taking the value $+\infty$, and a matrix $A \in \mathbb{R}^{n\times p}$. We consider the following minimization problem, which we refer to as the *primal* problem:

$$\min_{x\in\mathbb{R}^p} \; h(x) + f(Ax). \tag{1}$$

Throughout this paper, we make the following assumptions regarding the problem:

- *f is Lipschitz-continuous and finite on* $\mathbb{R}^n$, i.e., there exists a constant B such that for all $x, y \in \mathbb{R}^n$, $|f(x) - f(y)| \leqslant B\|x - y\|$, where $\|\cdot\|$ denotes the Euclidean norm. Note that this implies that the domain of the Fenchel conjugate f^* is bounded. We denote by C the bounded domain of f^*. Thus, for all $z \in \mathbb{R}^n$, $f(z) = \max_{y\in C} y^\top z - f^*(y)$. In many situations, C is also closed but this is not always the case (in particular, when $f^*(y)$ tends to infinity when y tends to the boundary of C).
  Note that the boundedness of the domain of f^* is crucial and allows for simpler proof techniques with explicit constants (see a generalization in [10]).
- *h is lower-semicontinuous and μ-strongly convex on* $\mathbb{R}^p$. This implies that h^* is defined on $\mathbb{R}^p$, differentiable with $(1/\mu)$-Lipschitz continuous gradient [8, 28]. Note that the domain K of h may be strictly included in $\mathbb{R}^p$.

Moreover, we assume that the following quantities may be computed efficiently:

- *Subgradient of f*: for any $z \in \mathbb{R}^n$, a subgradient of f is any maximizer y of $\max_{y\in C} y^\top z - f^*(y)$.
- *Gradient of h^**: for any $z \in \mathbb{R}^p$, $(h^*)'(z)$ may be computed and is equal to the unique maximizer x of $\max_{x\in\mathbb{R}^p} x^\top z - h(x)$.

The values of the functions f, h, f^* and h^* will be useful to compute duality gaps but are not needed to run the algorithms. As shown in Section 2, there are many examples of pairs of functions with the computational constraints described above. If other operations are possible, in particular $\max_{y\in C} y^\top z - f^*(y) - \frac{\varepsilon}{2}\|y\|^2$, then proximal methods [5, 26] applied to the dual problem converge at rate $O(1/t^2)$. If f and h are smooth, then gradient methods (accelerated [25, Section 2.2] or not) have linear convergence rates.

We denote by $g_{\text{primal}}(x) = h(x) + f(Ax)$ the primal objective in Eq. (1). It is the sum of a Lipschitz-continuous convex function and a strongly convex function, potentially on a restricted domain K. It is thus well adapted to the subgradient method [29].

We have the following primal/dual relationships (obtained from Fenchel duality [8]):

$$\begin{aligned}\min_{x\in\mathbb{R}^p} h(x)+f(Ax) &= \min_{x\in\mathbb{R}^p}\max_{y\in C} h(x)+y^\top(Ax)-f^*(y)\\ &= \max_{y\in C}\left\{\min_{x\in\mathbb{R}^p} h(x)+x^\top A^\top y\right\}-f^*(y)\\ &= \max_{y\in C} -h^*(-A^\top y)-f^*(y).\end{aligned}$$

This leads to the *dual* maximization problem:

$$\max_{y\in C} -h^*(-A^\top y)-f^*(y). \tag{2}$$

We denote by $g_{\text{dual}}(y) = -h^*(-A^\top y)-f^*(y)$ the dual objective. It has a smooth part $-h^*(-A^\top y)$ defined on $\mathbb{R}^n$ and a potentially non-smooth part $-f^*(y)$, and the problem is restricted onto a *bounded* set C. When f^* is linear (and more generally smooth) on its support, then we are exactly in the situation where conditional gradient algorithms may be used [16, 13].

Given a pair of primal-dual candidates $(x,y)\in K\times C$, we denote by $\text{gap}(x,y)$ the duality gap:

$$\text{gap}(x,y) = g_{\text{primal}}(x)-g_{\text{dual}}(y) = \big[h(x)+h^*(-A^\top y)+y^\top Ax\big]+\big[f(Ax)+f^*(y)-y^\top Ax\big].$$

It is equal to zero if and only if (a) $(x,-A^\top y)$ is a Fenchel-dual pair for h and (b) (Ax,y) is a Fenchel-dual pair for f. This quantity serves as a certificate of optimality, as

$$\text{gap}(x,y) = \big[g_{\text{primal}}(x)-\min_{x'\in K} g_{\text{primal}}(x')\big]+\big[\max_{y'\in C} g_{\text{dual}}(y')-g_{\text{dual}}(y)\big].$$

The goal of this paper is to show that for certain problems (f^* linear and h quadratic), the subgradient method applied to the primal problem in Eq. (1) is equivalent to the conditional gradient applied to the dual problem in Eq. (2); when relaxing the assumptions above, this equivalence is then between mirror descent methods and generalized conditional gradient algorithms.

2 Examples

The non-smooth strongly convex optimization problem defined in Eq. (1) occurs in many applications in machine learning and signal processing, either because they are formulated directly in this format, or their dual in Eq. (2) is (i.e., the original problem is the minimization of a smooth function over a compact set).

2.1 Direct formulations

Typical cases for h (often the regularizer in machine learning and signal processing) are the following:

– *Squared Euclidean norm*: $h(x) = \frac{\mu}{2}\|x\|^2$, which is μ-strongly convex.
– *Squared Euclidean norm with convex constraints*: $h(x) = \frac{\mu}{2}\|x\|^2 + I_K(x)$, with I_K the indicator function for K a closed convex set, which is μ-strongly convex.
– *Negative entropy*: $h(x) = \sum_{i=1}^n x_i \log x_i + I_K(x)$, where $K = \{x \in \mathbb{R}^n,\ x \geqslant 0,\ \sum_{i=1}^n x_i = 1\}$, which is 1-strongly convex. More generally, many barrier functions of convex sets may be used (see examples in [4, 9], in particular for problems on matrices).

Typical cases for f (often the data fitting terms in machine learning and signal processing) are functions of the form $f(z) = \frac{1}{n}\sum_{i=1}^n \ell_i(z_i)$:

– *Least-absolute-deviation*: $\ell_i(z_i) = |z_i - y_i|$, with $y_i \in \mathbb{R}$. Note that the square loss is not Lipschitz-continuous on $\mathbb{R}$ (although it is Lipschitz-continuous when restricted to a bounded set).
– *Logistic regression*: $\ell_i(z_i) = \log(1+\exp(-z_i y_i))$, with $y_i \in \{-1, 1\}$. Here f^* is not linear in its support, and f^* is not smooth, since it is a sum of negative entropies (and the second-order derivative is not bounded). This extends to any "log-sum-exp" functions which occur as a negative log-likelihood from the exponential family (see, e.g., [32] and references therein). Note that f is then smooth and proximal methods with an exponential convergence rate may be used (which correspond to a constant step size in the algorithms presented below, instead of a decaying step size) [26, 5].
– *Support vector machine*: $\ell_i(z_i) = \max\{1 - y_i z_i, 0\}$, with $y_i \in \{-1, 1\}$. Here f^* is linear on its domain (this is a situation where subgradient and conditional gradient methods are exactly equivalent). This extends to more general "max-margin" formulations [31, 30]: in these situations, a combinatorial object (such as a full chain, a graph, a matching or vertices of the hypercube) is estimated (rather than an element of $\{-1, 1\}$) and this leads to functions $z_i \mapsto \ell_i(z_i)$ whose Fenchel-conjugates are linear and have domains which are related to the polytopes associated to the linear programming relaxations of the corresponding combinatorial optimization problems. For these polytopes, often, only linear functions can be maximized, i.e., we can compute a subgradient of ℓ_i but typically nothing more.

Other examples may be found in signal processing; for example, total-variation denoising, where the loss is strongly convex but the regularizer is non-smooth [11], or submodular function minimization cast through separable optimization problems [2]. Moreover, many proximal operators for non-smooth regularizers are of this form, with $h(x) = \frac{1}{2}\|x - x_0\|^2$ and f is a norm (or more generally a gauge function).

2.2 Dual formulations

Another interesting set of examples for machine learning are more naturally described from the dual formulation in Eq. (2): given a smooth loss term $h^*(-A^\top y)$ (this could be least-squares or logistic regression), a typically non-smooth penalization or constraint is added,

often through a norm Ω. Thus, this corresponds to functions f^* of the form $f^*(y) = \varphi(\Omega(y))$, where φ is a convex non-decreasing function (f^* is then convex).

Our main assumption is that a subgradient of f may be easily computed. This is equivalent to being able to maximize functions of the form $z^\top y - f^*(y) = z^\top y - \varphi(\Omega(y))$ for $z \in \mathbb{R}^n$. If one can compute the dual norm of z, $\Omega^*(z) = \max_{\Omega(y)\leqslant 1} z^\top y$, and in particular a maximizer y in the unit-ball of Ω, then one can compute simply the subgradient of f. Only being able to compute the dual norm efficiently is a common situation in machine learning and signal processing, for example, for structured regularizers based on submodularity [2], all atomic norms [12], and norms based on matrix decompositions [1]. See additional examples in [19].

Our assumption regarding the compact domain of f^* translates to the assumption that φ has compact domain. This includes indicator functions $\varphi = I_{[0,\omega_0]}$ which corresponds to the constraint $\Omega(y) \leqslant \omega_0$. We may also consider $\varphi(\omega) = \lambda\omega + I_{[0,\omega_0]}(\omega)$, which corresponds to jointly penalizing and constraining the norm; in practice, ω_0 may be chosen so that the constraint $\Omega(y) \leqslant \omega_0$ is not active at the optimum and we get the solution of the penalized problem $\max_{y\in\mathbb{R}^n} -h^*(-A^\top y) - \lambda\Omega(y)$. See [17, 34, 1] for alternative approaches.

3 Mirror descent for strongly convex problems

We first assume that the function h is *essentially smooth* (i.e., differentiable at any point in the interior of K, and so that the norm of gradients converges to $+\infty$ when approaching the boundary of K); then h' is a bijection from $\mathrm{int}(K)$ to $\mathbb{R}^p$, where K is the domain of h (see, e.g., [28, 18]). We consider the Bregman divergence

$$D(x_1, x_2) = h(x_1) - h(x_2) - (x_1 - x_2)^\top h'(x_2).$$

It is always defined on $K \times \mathrm{int}(K)$, and is nonnegative. If $x_1, x_2 \in \mathrm{int}(K)$, then $D(x_1, x_2) = 0$ if and only if $x_1 = x_2$. Moreover, since h is assumed μ-strongly convex, we have $D(x_1, x_2) \geqslant \frac{\mu}{2}\|x_1 - x_2\|^2$. See more details in [4]. For example, when $h(x) = \frac{\mu}{2}\|x\|^2$, we have $D(x_1, x_2) = \frac{\mu}{2}\|x_1 - x_2\|^2$.

Subgradient descent for square Bregman divergence We first consider the common situation where $h(x) = \frac{\mu}{2}\|x\|^2$; the primal problem then becomes:

$$\min_{x\in K} f(Ax) + \frac{\mu}{2}\|x\|^2.$$

The projected subgradient method starts from any $x_0 \in \mathbb{R}^p$, and iterates the following recursion:

$$x_t = x_{t-1} - \frac{\rho_t}{\mu}\big[A^\top f'(Ax_{t-1}) + \mu x_{t-1}\big],$$

where $\bar{y}_{t-1} = f'(Ax_{t-1})$ is any subgradient of f at Ax_{t-1}. The step size is $\dfrac{\rho_t}{\mu}$.

The recursion may be rewritten as

$$\mu x_t = \mu x_{t-1} - \rho_t\big[A^\top f'(Ax_{t-1}) + \mu x_{t-1}\big],$$

which is equivalent to x_t being the unique minimizer of

$$(x - x_{t-1})^\top \big[A^\top \bar{y}_{t-1} + \mu x_{t-1}\big] + \frac{\mu}{2\rho_t}\|x - x_{t-1}\|^2, \tag{3}$$

which is the traditional proximal step, with step size ρ_t/μ.

Mirror descent We may interpret the last formulation in Eq. (3) for the square regularizer $h(x) = \frac{\mu}{2}\|x\|^2$ as the minimization of

$$(x - x_{t-1})^\top g'_{\text{primal}}(x_{t-1}) + \frac{1}{\rho_t} D(x, x_{t-1}),$$

with solution defined through (note that h' is a bijection from $\text{int}(K)$ to $\mathbb{R}^p$):

$$\begin{aligned} h'(x_t) &= h'(x_{t-1}) - \rho_t\big[A^\top f'(Ax_{t-1}) + h'(x_{t-1})\big] \\ &= (1-\rho_t)h'(x_{t-1}) - \rho_t A^\top f'(Ax_{t-1}). \end{aligned}$$

This leads to the following definition of the mirror descent recursion:

$$\left\{ \begin{array}{rcl} \bar{y}_{t-1} & \in & \arg\max\limits_{y \in C} \; y^\top A x_{t-1} - f^*(y), \\ x_t & = & \arg\min\limits_{x \in \mathbb{R}^p} \; h(x) - (1-\rho_t) x^\top h'(x_{t-1}) + \rho_t x^\top A^\top \bar{y}_{t-1}. \end{array} \right. \tag{4}$$

The following proposition proves the convergence of mirror descent in the strongly convex case with rate $O(1/t)$—previous results were considering the convex case, with convergence rate $O(1/\sqrt{t})$ [24, 4].

Proposition 1 (Convergence of mirror descent in the strongly convex case) *Assume that (a) f is Lipschitz-continuous and finite on $\mathbb{R}^p$, with C the domain of f^*, (b) h is essentially smooth and μ-strongly convex. Consider $\rho_t = 2/(t+1)$ and $R^2 = \max_{y,y' \in C} \|A^\top(y - y')\|^2$. Denoting by x_* the unique minimizer of g_{primal}, after t iterations of the mirror descent recursion of Eq. (4), we have:*

$$\begin{aligned} g\bigg(\frac{2}{t(t+1)}\sum_{u=1}^t u x_{u-1}\bigg) - g_{\text{primal}}(x_*) &\leqslant \frac{R^2}{\mu(t+1)}, \\ \min_{u \in \{0,\dots,t-1\}} \Big\{ g_{\text{primal}}(x_u) - g_{\text{primal}}(x_*) \Big\} &\leqslant \frac{R^2}{\mu(t+1)}, \\ D(x_*, x_t) &\leqslant \frac{R^2}{\mu(t+1)}. \end{aligned}$$

Proof We follow the proof of [4] and adapt it to the strongly convex case. We have, by reordering terms and using the optimality condition $h'(x_t) = h'(x_{t-1}) - \rho_t\big[A^\top f'(Ax_{t-1}) +$

$h'(x_{t-1})\big]$:

$$\begin{aligned}
& D(x_*, x_t) - D(x_*, x_{t-1}) \\
= \;& h(x_{t-1}) - h(x_t) - (x_* - x_t)^\top h'(x_t) + (x_* - x_{t-1})^\top h'(x_{t-1}) \\
= \;& h(x_{t-1}) - h(x_t) - (x_* - x_t)^\top \big[(1-\rho_t) h'(x_{t-1}) - \rho_t A^\top f'(Ax_{t-1})\big] \\
& \qquad + (x_* - x_{t-1})^\top h'(x_{t-1}) \\
= \;& h(x_{t-1}) - h(x_t) - (x_{t-1} - x_t)^\top h'(x_{t-1}) + \rho_t (x_* - x_t)^\top g'_{\text{primal}}(x_{t-1}) \\
= \;& \big[- D(x_t, x_{t-1}) + \rho_t (x_{t-1} - x_t)^\top g'_{\text{primal}}(x_{t-1})\big] \\
& \qquad + \big[\rho_t (x_* - x_{t-1})^\top g'_{\text{primal}}(x_{t-1})\big].
\end{aligned} \tag{5}$$

In order to upper-bound the two terms in Eq. (5), we first consider the following bound (obtained by convexity of f and the definition of D):

$$f(Ax_*) + h(x_*) \geqslant f(Ax_{t-1}) + h(x_{t-1}) + (x_* - x_{t-1})^\top [A^\top \bar{y}_{t-1} + h'(x_{t-1})] + D(x_*, x_{t-1}),$$

which may be rewritten as:

$$g_{\text{primal}}(x_{t-1}) - g_{\text{primal}}(x_*) \leqslant -D(x_*, x_{t-1}) + (x_{t-1} - x_*)^\top g'_{\text{primal}}(x_{t-1}),$$

which implies

$$\rho_t (x_* - x_{t-1})^\top g'_{\text{primal}}(x_{t-1}) \leqslant -\rho_t D(x_*, x_{t-1}) - \rho_t \big[g_{\text{primal}}(x_{t-1}) - g_{\text{primal}}(x_*)\big]. \tag{6}$$

Moreover, by definition of x_t,

$$-D(x_t, x_{t-1}) + \rho_t (x_{t-1} - x_t)^\top g'_{\text{primal}}(x_{t-1}) = \max_{x \in \mathbb{R}^p} -D(x, x_{t-1}) + \rho_t (x_{t-1} - x)^\top z = \varphi(z),$$

with $z = \rho_t g'_{\text{primal}}(x_{t-1})$. The function $x \mapsto D(x, x_{t-1})$ is μ-strongly convex, and its Fenchel conjugate is thus $(1/\mu)$-smooth. This implies that φ is $(1/\mu)$-smooth. Since $\varphi(0) = 0$ and $\varphi'(0) = 0$, $\varphi(z) \leqslant \frac{1}{2\mu}\|z\|^2$. Moreover, $z = \rho_t \big[A^\top f'(Ax_{t-1}) + h(x_{t-1})\big]$. Since $h'(x_{t-1}) \in -A^\top C$ (because $h'(x_{t-1})$ is a convex combination of such elements), then $\|A^\top f'(Ax_{t-1}) + h(x_{t-1})\|^2 \leqslant R^2 = \max_{y_1, y_2 \in C} \|A^\top (y_1 - y_2)\|^2 = \operatorname{diam}(A^\top C)^2$.

Overall, combining Eq. (6) and $\varphi(z) \leqslant \frac{R^2 \rho_t^2}{2\mu}$ into Eq. (5), this implies that

$$D(x_*, x_t) - D(x_*, x_{t-1}) \leqslant \frac{\rho_t^2}{2\mu} R^2 - \rho_t D(x_*, x_{t-1}) - \rho_t \big[g_{\text{primal}}(x_{t-1}) - g_{\text{primal}}(x_*)\big],$$

that is,

$$g_{\text{primal}}(x_{t-1}) - g_{\text{primal}}(x_*) \leqslant \frac{\rho_t R^2}{2\mu} + (\rho_t^{-1} - 1) D(x_*, x_{t-1}) - \rho_t^{-1} D(x_*, x_t).$$

With $\rho_t = \dfrac{2}{t+1}$, we obtain

$$t\big[g_{\text{primal}}(x_{t-1}) - g_{\text{primal}}(x_*)\big] \;\leqslant\; \frac{R^2 t}{\mu(t+1)} + \frac{(t-1)t}{2} D(x_*, x_{t-1}) - \frac{t(t+1)}{2} D(x_*, x_t).$$

Thus, by summing from $u = 1$ to $u = t$, we obtain

$$\sum_{u=1}^{t} u\big[g_{\text{primal}}(x_{u-1}) - g_{\text{primal}}(x_*)\big] \leqslant \frac{R^2}{\mu} t - \frac{t(t+1)}{2} D(x_*, x_t),$$

that is,

$$D(x_*, x_t) + \frac{2}{t(t+1)} \sum_{u=1}^{t} u\big[g_{\text{primal}}(x_{u-1}) - g_{\text{primal}}(x_*)\big] \leqslant \frac{R^2}{\mu(t+1)}.$$

This implies that $D(x_*, x_t) \leqslant \dfrac{R^2}{\mu(t+1)}$, i.e., the iterates converges. Moreover, using the convexity of g,

$$g\bigg(\frac{2}{t(t+1)} \sum_{u=1}^{t} u x_{u-1}\bigg) - g_{\text{primal}}(x_*) \leqslant \frac{2}{t(t+1)} \sum_{u=1}^{t} u\big[g_{\text{primal}}(x_{u-1}) - g_{\text{primal}}(x_*)\big] \leqslant \frac{R^2}{\mu(t+1)},$$

i.e., the objective functions at an averaged iterate converges, and

$$\min_{u \in \{0,\dots,t-1\}} g_{\text{primal}}(x_u) - g_{\text{primal}}(x_*) \leqslant \frac{R^2}{\mu(t+1)},$$

i.e., one of the iterates has an objective that converges. ■

Averaging Note that with the step size $\rho_t = \dfrac{2}{t+1}$, we have

$$h'(x_t) = \frac{t-1}{t+1} h'(x_{t-1}) - \frac{2}{t+1} A^\top f'(A x_{t-1}),$$

which implies

$$t(t+1)h'(x_t) = (t-1)t h'(x_{t-1}) - 2t A^\top f'(A x_{t-1}).$$

By summing these equalities, we obtain $t(t+1)h'(x_t) = -2\sum_{u=1}^{t} u A^\top f'(A x_{u-1})$, i.e.,

$$h'(x_t) = \frac{2}{t(t+1)} \sum_{u=1}^{t} u\big[-A^\top f'(A x_{u-1})\big],$$

that is, $h'(x_t)$ is a weighted average of subgradients (with more weights on later iterates).

For $\rho_t = 1/t$, then, we the same techniques, we would obtain a convergence rate proportional to $\frac{R^2}{\mu t} \log t$ for the average iterate $\frac{1}{t}\sum_{u=1}^{t} x_{u-1}$, thus with an additional $\log t$ factor (see a similar situation in the stochastic case in [20]). We would then have $h'(x_t) = \frac{1}{t}\sum_{u=1}^{t}\big[-A^\top f'(A x_{u-1})\big]$, and this is exactly a form dual averaging method [27], which also comes with primal-dual guarantees.

Generalization to h non-smooth The previous result does not require h to be essentially smooth, i.e., it may be applied to $h(x) = \frac{\mu}{2}\|x\|^2 + I_K(x)$ where K is a closed convex set strictly included in $\mathbb{R}^p$. In the mirror descent recursion,

$$\left\{ \begin{array}{rcl} \bar{y}_{t-1} & \in & \arg\max\limits_{y \in C} \; y^\top A x_{t-1} - f^*(y), \\ x_t & = & \arg\min\limits_{x \in \mathbb{R}^p} \; h(x) - (1-\rho_t) x^\top h'(x_{t-1}) + \rho_t x^\top A^\top \bar{y}_{t-1}, \end{array} \right.$$

there may then be multiple choices for $h'(x_{t-1})$. If we choose for $h'(x_{t-1})$ at iteration t, the subgradient of h obtained at the previous iteration, i.e., such that $h'(x_{t-1}) = (1-\rho_{t-1})h'(x_{t-2}) - \rho_{t-1}A^\top \bar{y}_{t-2}$, then the proof of Prop. 1 above holds.

Note that when $h(x) = \frac{\mu}{2}\|x\|^2 + I_K(x)$, the algorithm above is *not* equivalent to classical projected gradient descent. Indeed, the classical algorithm has the iteration

$$x_t = \Pi_K\Big(x_{t-1} - \frac{1}{\mu}\rho_t\big[\mu x_{t-1} + A^\top f'(Ax_{t-1})\big]\Big) = \Pi_K\Big((1-\rho_t)x_{t-1} + \rho_t\big[-\frac{1}{\mu}A^\top f'(Ax_{t-1})\big]\Big),$$

and corresponds to the choice $h'(x_{t-1}) = \mu x_{t-1}$ in the mirror descent recursion, which, when x_{t-1} is on the boundary of K, is not the choice that we need for the equivalence in Section 4.

However, when h is assumed to be differentiable on its closed domain K, then the bound of Prop. 1 still holds because the optimality condition $h'(x_t) = h'(x_{t-1}) - \rho_t\big[A^\top f'(Ax_{t-1}) + h'(x_{t-1})\big]$ may now be replaced by $(x - x_t)^\top\Big(h'(x_t) - h'(x_{t-1}) + \rho_t\big[A^\top f'(Ax_{t-1}) + h'(x_{t-1})\big]\Big) \geqslant 0$ for all $x \in K$, which also allows to get to Eq. (5) in the proof of Prop. 1.

4 Conditional gradient method and extensions

In this section, we first review the classical conditional gradient algorithm, which corresponds to the extra assumption that f^* is linear in its domain.

Conditional gradient method Given a maximization problem of the following form (i.e., where f^* is linear on its domain, or equal to zero by a simple change of variable):

$$\max_{y \in C} -h^*(-A^\top y),$$

the conditional gradient algorithm consists in the following iteration (note that below $Ax_{t-1} = A(h^*)'(-A^\top y_{t-1})$ is the gradient of the objective function and that we are maximizing the first-order Taylor expansion to obtain a candidate $\bar{y}_{t-1}$ towards which we make a small step):

$$\begin{array}{rcl} x_{t-1} & = & \arg\min\limits_{x \in \mathbb{R}^p} h(x) + x^\top A^\top y_{t-1} \\ \bar{y}_{t-1} & \in & \arg\max\limits_{y \in C} y^\top A x_{t-1} \\ y_t & = & (1-\rho_t) y_{t-1} + \rho_t \bar{y}_{t-1}. \end{array}$$

It corresponds to a linearization of $-h^*(-A^\top y)$ and its maximization over the bounded convex set C. As we show later, the choice of ρ_t may be done in different ways, through a fixed step size of by (approximate) line search.

Generalization Following [10], the conditional gradient method can be generalized to problems of the form

$$\max_{y \in C} -h^*(-A^\top y) - f^*(y),$$

with the following iteration:

$$\left\{ \begin{array}{rcl} x_{t-1} & = & \arg\min_{x \in \mathbb{R}^p} h(x) + x^\top A^\top y_{t-1} = (h^*)'(-A^\top y_{t-1}) \\ \bar{y}_{t-1} & \in & \arg\max_{y \in C} y^\top A x_{t-1} - f^*(y) \\ y_t & = & (1-\rho_t) y_{t-1} + \rho_t \bar{y}_{t-1}. \end{array} \right. \tag{7}$$

The previous algorithm may be interpreted as follows: (a) perform a first-order Taylor expansion of the smooth part $-h^*(-A^\top y)$, while leaving the other part $-f^*(y)$ intact, (b) minimize the approximation, and (c) perform a small step towards the maximizer. Note the similarity (and dissimilarity) with proximal methods which would add a proximal term proportional to $\|y - y_{t-1}\|^2$, leading to faster convergences, but with the extra requirement of solving the proximal step [26, 5].

Note that here y_t may be expressed as a *convex* combination of all $\bar{y}_{u-1}$, $u \in \{1, \dots, t\}$:

$$y_t = \sum_{u=1}^{t} \left(\rho_u \prod_{s=u+1}^{t} (1-\rho_s) \right) \bar{y}_{u-1},$$

and that when we chose $\rho_t = 2/(t+1)$, it simplifies to:

$$y_t = \frac{2}{t(t+1)} \sum_{u=1}^{t} u \bar{y}_{u-1}.$$

When h is essentially smooth (and thus h^* is essentially strictly convex), it can be reformulated with $h'(x_t) = -A^\top y_t$ as follows:

$$\begin{array}{rcl} h'(x_t) & = & (1-\rho_t) h'(x_{t-1}) - \rho_t A^\top \arg\max_{y \in C} \left\{ y^\top A x_{t-1} - f^*(y) \right\}, \\ & = & (1-\rho_t) h'(x_{t-1}) - \rho_t A^\top f'(A x_{t-1}), \end{array}$$

which is exactly the mirror descent algorithm described in Eq. (4). This leads to the following proposition:

Proposition 2 (Equivalence between mirror descent and generalized conditional gradient) *Assume that (a) f is Lipschitz-continuous and finite on $\mathbb{R}^p$, with C the domain of f^*, (b) h is μ-strongly convex and essentially smooth. The mirror descent recursion in Eq. (4), started from $x_0 = (h^*)'(-A^\top y_0)$, is equivalent to the generalized conditional gradient recursion in Eq. (7), started from $y_0 \in C$.*

When h is not essentially smooth, then with a particular choice of subgradient (see end of Section 3), the two algorithms are also equivalent. We now provide convergence proofs for the two versions (with adaptive and non-adaptive step sizes); similar rates may be obtained without the boundedness assumptions [10], but our results provide explicit constants and primal-dual guarantees. We first have the following convergence proof for generalized conditional gradient with no line search (the proof of dual convergence uses standard arguments from [13, 15], while the convergence of gaps is due to [19] for the regular conditional gradient):

Proposition 3 (Convergence of extended conditional gradient - no line search) *Assume that (a) f is Lipschitz-continuous and finite on $\mathbb{R}^p$, with C the domain of f^*, (b) h is μ-strongly convex. Consider $\rho_t = 2/(t+1)$ and $R^2 = \max_{y,y' \in C} \|A^\top (y-y')\|^2$. Denoting by y_* any maximizer of g_{dual} on C, after t iterations of the generalized conditional gradient recursion of Eq. (7), we have:*

$$\begin{aligned} g_{\text{dual}}(y_*) - g_{\text{dual}}(y_t) &\leqslant \frac{2R^2}{\mu(t+1)}, \\ \min_{u \in \{0,\dots,t-1\}} \text{gap}(x_t, y_t) &\leqslant \frac{8R^2}{\mu(t+1)}. \end{aligned}$$

Proof We have (using convexity of f^* and $\left(\frac{1}{\mu}\right)$-smoothness of h^*):

$$\begin{aligned} & g_{\text{dual}}(y_t) \\ = \; & -h^*(-A^\top y_t) - f^*(y_t) \\ \geqslant \; & \left[- h^*(-A^\top y_{t-1}) + (y_t - y_{t-1})^\top A x_{t-1} - \frac{R^2 \rho_t^2}{2\mu} \right] - \left[(1-\rho_t) f^*(y_{t-1}) + \rho_t f^*(\bar{y}_{t-1}) \right] \\ = \; & -h^*(-A^\top y_{t-1}) + \rho_t (\bar{y}_{t-1} - y_{t-1})^\top A x_{t-1} - \frac{R^2 \rho_t^2}{2\mu} - (1-\rho_t) f^*(y_{t-1}) - \rho_t f^*(\bar{y}_{t-1}) \\ = \; & g_{\text{dual}}(y_{t-1}) + \rho_t (\bar{y}_{t-1} - y_{t-1})^\top A x_{t-1} - \frac{R^2 \rho_t^2}{2\mu} + \rho_t f^*(y_{t-1}) - \rho_t f^*(\bar{y}_{t-1}) \\ = \; & g_{\text{dual}}(y_{t-1}) - \frac{R^2 \rho_t^2}{2\mu} + \rho_t \left[f^*(y_{t-1}) - f^*(\bar{y}_{t-1}) + (\bar{y}_{t-1} - y_{t-1})^\top A x_{t-1} \right] \\ = \; & g_{\text{dual}}(y_{t-1}) - \frac{R^2 \rho_t^2}{2\mu} + \rho_t \left[f^*(y_{t-1}) - y_{t-1}^\top A x_{t-1} - (f^*(\bar{y}_{t-1}) - \bar{y}_{t-1}^\top A x_{t-1}) \right]. \end{aligned}$$

Note that by definition of $\bar{y}_{t-1}$, we have (by equality in Fenchel-Young inequality)

$$-f^*(\bar{y}_{t-1}) + \bar{y}_{t-1}^\top A x_{t-1} = f(A x_{t-1}),$$

and $h^*(-A^\top y_{t-1}) + h(x_{t-1}) + x_{t-1}^\top A^\top y_{t-1} = 0$, and thus

$$f^*(y_{t-1}) - y_{t-1}^\top A x_{t-1} - (f^*(\bar{y}_{t-1}) - \bar{y}_{t-1}^\top A x_{t-1}) = g_{\text{primal}}(x_{t-1}) - g_{\text{dual}}(y_{t-1}) = \text{gap}(x_{t-1}, y_{t-1}).$$

We thus obtain, for any $\rho_t \in [0,1]$:

$$g_{\text{dual}}(y_t) - g_{\text{dual}}(y_*) \geqslant g_{\text{dual}}(y_{t-1}) - g_{\text{dual}}(y_*) + \rho_t \text{gap}(x_{t-1}, y_{t-1}) - \frac{R^2 \rho_t^2}{2\mu},$$

which is the classical equation from the conditional gradient algorithm [15, 14, 19], which we can analyze through Lemma 1 (see end of this section), leading to the desired result. ■

The following proposition shows a result similar to the proposition above, but for the adaptive algorithm that considers optimizing the value ρ_t at each iteration.

Proposition 4 (Convergence of extended conditional gradient - with line search) *Assume that (a) f is Lipschitz-continuous and finite on $\mathbb{R}^p$, with C the domain of f^*, (b) h is μ-strongly convex. Consider $\rho_t = \min\{\frac{\mu}{R^2}\text{gap}(x_{t-1}, y_{t-1}), 1\}$ and $R^2 = \max_{y,y' \in C} \|A^\top(y - y')\|^2$. Denoting by y_* any maximizer of g_{dual} on C, after t iterations of the generalized conditional gradient recursion of Eq. (7), we have:*

$$\begin{aligned} g_{\text{dual}}(y_*) - g_{\text{dual}}(y_t) &\leqslant \frac{2R^2}{\mu(t+3)}, \\ \min_{u \in \{0,\dots,t-1\}} \text{gap}(x_t, y_t) &\leqslant \frac{2R^2}{\mu(t+3)}. \end{aligned}$$

Proof The proof is essentially the same as one from the previous proposition, with a different application of Lemma 1 (see below). ■

The following technical lemma is used in the previous proofs to obtain the various convergence rates.

Lemma 1 *Assume that we have three sequences $(u_t)_{t \geqslant 0}$, $(v_t)_{t \geqslant 0}$, and $(\rho_t)_{t \geqslant 0}$, and a positive constant A such that*

$$\begin{aligned} &\forall t \geqslant 0,\ \rho_t \in [0, 1] \\ &\forall t \geqslant 0,\ 0 \leqslant u_t \leqslant v_t \\ &\forall t \geqslant 1,\ u_t \leqslant u_{t-1} - \rho_t v_{t-1} + \frac{A}{2}\rho_t^2. \end{aligned}$$

- *If $\rho_t = 2/(t+1)$, then $u_t \leqslant \frac{2A}{t+1}$ and for all $t \geqslant 1$, there exists at least one $k \in \{\lfloor t/2 \rfloor, \dots, t\}$ such that $v_k \leqslant \frac{8A}{t+1}$.*
- *If $\rho_t = \arg\min_{\rho_t \in [0,1]} -\rho_t v_{t-1} + \frac{A}{2}\rho_t^2 = \min\{v_{t-1}/A, 1\}$, then $u_t \leqslant \frac{2A}{t+3}$ and for all $t \geqslant 2$, there exists at least one $k \in \{\lfloor t/2 \rfloor - 1, \dots, t\}$ such that $v_k \leqslant \frac{2A}{t+3}$.*

Proof In the first case (non-adaptive sequence ρ_t), we have $\rho_0 = 1$ and $u_t \leqslant (1 - \rho_t)u_{t-1} + \frac{A}{2}\rho_t^2$, leading to

$$u_t \leqslant \frac{A}{2} \sum_{u=1}^{t} \prod_{s=u+1}^{t} (1 - \rho_s)\rho_u^2.$$

For $\rho_t = \frac{2}{t+1}$, this leads to

$$u_t \leqslant \frac{A}{2} \sum_{u=1}^{t} \prod_{s=u+1}^{t} \frac{s-1}{s+1} =\leqslant \frac{A}{2} \sum_{u=1}^{t} \frac{u(u+1)}{t(t+1)} \frac{4}{(u+1)^2} \leqslant \frac{2A}{t+1}.$$

Moreover, for any $k < j$, by summing $u_t \leqslant u_{t-1} - \rho_t v_{t-1} + \frac{A}{2}\rho_t^2$ for $t \in \{k+1, \dots, j\}$, we get

$$u_j \leqslant u_k - \sum_{t=k+1}^{j} \rho_t v_{t-1} + \frac{A}{2} \sum_{t=k+1}^{j} \rho_t^2.$$

Thus, if we assume that $v_{t-1} \geqslant \beta$ for all $t \in \{k+1, \dots, j\}$, then

$$\begin{aligned}
\beta \sum_{t=k+1}^{j} \rho_t \leqslant \sum_{t=k+1}^{j} \rho_t v_{t-1} &\leqslant \frac{2A}{k+1} + 2A \sum_{t=k+1}^{j} \frac{1}{(t+1)^2} \\
&\leqslant \frac{2A}{k+1} + 2A \sum_{t=k+1}^{j} \frac{1}{t(t+1)} \\
&= \frac{2A}{k+1} + 2A \sum_{t=k+1}^{j} \big[\frac{1}{t} - \frac{1}{t+1}\big] \leqslant \frac{4A}{k+1}.
\end{aligned}$$

Moreover, $\sum_{t=k+1}^{j} \rho_t = 2\sum_{t=k+1}^{j} \frac{1}{t+1} \geqslant 2\frac{j-k}{j+1}$. Thus

$$\beta \leqslant \frac{2A}{k+1}\frac{j+1}{j-k}.$$

Using $j = t+1$ and $k = \lfloor t/2 \rfloor - 1$, we obtain that $\beta \leqslant \frac{8A}{t+1}$ (this can be done by considering the two cases t even and t odd) and thus $\max_{u \in \{\lfloor t/2 \rfloor, \dots, t\}} v_u \leqslant \frac{8A}{t+1}$.

We now consider the line search case:

- If $v_{t-1} \leqslant A$, then $\rho_t = \frac{v_{t-1}}{A}$, and we obtain $u_t \leqslant u_{t-1} - \frac{v_{t-1}^2}{2A}$.
- If $v_{t-1} \geqslant A$, then $\rho_t = 1$, and we obtain $u_t \leqslant u_{t-1} - v_{t-1} + \frac{A}{2} \leqslant u_{t-1} - \frac{v_{t-1}}{2}$.

Putting all this together, we get $u_t \leqslant u_{t-1} - \frac{1}{2}\min\{v_{t-1}, v_{t-1}^2/A\}$. This implies that (u_t) is a decreasing sequence. Moreover, $u_1 \leqslant \frac{A}{2}$ (because selecting $\rho_1 = 1$ leads to this value), thus, $u_1 \leqslant \min\{u_0, A/2\} \leqslant A$. We then obtain for all $t > 1$, $u_t \leqslant u_{t-1} - \frac{1}{2A}u_{t-1}^2$. From which we deduce, $u_{t-1}^{-1} \leqslant u_t^{-1} - \frac{1}{2A}$. We can now sum these inequalities to get $u_1^{-1} \leqslant u_t^{-1} - \frac{t-1}{2A}$, that is,

$$u_t \leqslant \frac{1}{u_1^{-1} + \frac{t-1}{2A}} \leqslant \frac{1}{\max\{u_0^{-1}, 2/A\} + \frac{t-1}{2A}} \leqslant \frac{2A}{t+3}.$$

Moreover, if we assume that all $v_{t-1} \geqslant \beta$ for $t \in \{k+1, \dots, j\}$, following the same reasoning as above, and using the inequality $u_t \leqslant u_{t-1} - \frac{1}{2}\min\{v_{t-1}, v_{t-1}^2/A\}$, we obtain

$$\min\{\beta, \beta^2/A\}(j-k) \leqslant \frac{A}{k+3}.$$

Using $j = t+1$ and $k = \lfloor t/2 \rfloor - 1$, we have $(k+3)(j-k) > \frac{1}{4}(t+3)^2$ (which can be checked by considering the two cases t even and t odd). Thus, we must have $\beta \leqslant A$ (otherwise we obtain $\beta \leqslant 4A/(t+3)^2$, which is a contradiction with $\beta \geqslant A$), and thus $\beta^2 \leqslant 4A^2/(t+3)^2$, which leads to the desired result. ■

5 Discussion

The equivalence shown in Prop. 2 has several interesting consequences and leads to several additional related questions:

- **Primal-dual guarantees**: Having a primal-dual interpretation directly leads to primal-dual certificates, with a gap that converges at the same rate proportional to $\frac{R^2}{\mu t}$ (see [19, 20] for similar results for the regular conditional gradient method). These certificates may first be taken to be the pair (x_t, y_t), in which case, we have shown that after t iterations, at least one of the previous iterates has the guarantee. Alternatively, for the fixed step-size $\rho_t = \frac{2}{t+1}$, we can use the same dual candidate $y_t = \frac{2}{t(t+1)} \sum_{u=1}^{t} u \bar{y}_{u-1}$ (which can thus also be expressed as an average of subgradients) and averaged primal iterate $\frac{2}{t(t+1)} \sum_{u=1}^{t} u x_{u-1}$. Thus, the two weighted averages of subgradients lead to primal-dual certificates.
- **Line-search for mirror descent**: Prop. 4 provides a form of line search for mirror descent (i.e., an adaptive step size). Note the similarity with Polyak's rule which applies to the non-strongly convex case (see, e.g., [6]).
- **Absence of logarithmic terms**: Note that we have considered a step-size of $\frac{2}{t+1}$, which avoids a logarithmic term of the form $\log t$ in all bounds (which would be the case for $\rho_t = \frac{1}{t}$). This also applies to the stochastic case [21].
- **Properties of iterates**: While we have focused primarily on the convergence rates of the iterates and their objective values, recent work has shown that the iterates themselves could have interesting distributional properties [33, 3], which would be worth further investigation.
- **Stochastic approximation and online learning**: There are potentially other exchanges between primal/dual formulations, in particular in the stochastic setting (see, e.g., [20]).
- **Simplicial methods and cutting-planes**: The duality between subgradient and conditional gradient may be extended to algorithms with iterations that are more expensive. For example, simplicial methods in the dual are equivalent to cutting-planes methods in the primal (see, e.g., [7, 20] and [2, Chapter 7]).

Acknowledgements

This work was partially supported by the European Research Council (SIERRA Project). The author would like to thank Simon Lacoste-Julien, Martin Jaggi, Mark Schmidt and Zaid Harchaoui for discussions related to convex optimization and conditional gradient algorithms.

References

[1] F. Bach. Convex relaxations of structured matrix factorizations. Technical Report 00861118, HAL, 2013.

[2] F. Bach. Learning with submodular functions: A convex optimization perspective. Technical Report 1111.6453-v2, Arxiv, 2013.

[3] F. Bach, S. Lacoste-Julien, and G. Obozinski. On the equivalence between herding and conditional gradient algorithms. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

[4] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[5] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[6] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific Belmont, 1999.

[7] D. P. Bertsekas and H. Yu. A unifying polyhedral approximation framework for convex optimization. *SIAM Journal on Optimization*, 21(1):333–360, 2011.

[8] J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer, 2006.

[9] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[10] K. Bredies and D. A. Lorenz. Iterated hard shrinkage for minimization problems with sparsity constraints. *SIAM Journal on Scientific Computing*, 30(2):657–683, 2008.

[11] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[12] V. Chandrasekaran, B. Recht, P. A., and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[13] V. F. Dem'yanov and A. M. Rubinov. The minimization of a smooth convex functional on a convex set. *SIAM Journal on Control*, 5(2):280–294, 1967.

[14] J. C. Dunn. Convergence rates for conditional gradient sequences generated by implicit step length rules. *SIAM Journal on Control and Optimization*, 18:473–487, 1980.

[15] J. 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.

[16] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[17] Z. Harchaoui, A. Juditsky, and A. Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. Technical Report 1302.2325, arXiv, 2013.

[18] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms: Part 1: Fundamentals*, volume 1. Springer, 1996.

[19] M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2013.

[20] S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2013.

[21] S. Lacoste-Julien, M. Schmidt, and F. Bach. A simpler approach to obtaining an $o(1/t)$ convergence rate for the projected stochastic subgradient method. Technical Report 1212.2002, Arxiv, 2012.

[22] A. Nedic and D. Bertsekas. Convergence rate of incremental subgradient algorithms. In *Stochastic Optimization: Algorithms and Applications*, 2000.

[23] A. Nedic and A. Ozdaglar. Approximate primal solutions and rate analysis for dual subgradient methods. *SIAM Journal on Optimization*, 19(4), February 2009.

[24] A. S. Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in optimization*. John Wiley, 1983.

[25] Y. Nesterov. *Introductory lectures on convex optimization: a basic course*. Kluwer Academic Publishers, 2004.

[26] Y. Nesterov. Gradient methods for minimizing composite objective function. Technical report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, 2007.

[27] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

[28] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1997.

[29] N. Z. Shor, K. C. Kiwiel, and A. Ruszczynski. *Minimization methods for non-differentiable functions*. Springer-Verlag, 1985.

[30] B. Taskar, V. Chatalbashev, D. Koller, and C. Guestrin. Learning structured prediction models: A large margin approach. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2005.

[31] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6(2):1453–1484, 2006.

[32] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

[33] M. Welling. Herding dynamical weights to learn. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

[34] X. Zhang, D. Schuurmans, and Y. Yu. Accelerated training for matrix-norm regularization: A boosting approach. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

A Non-strongly convex case

In this appendix, we consider the situation where the primal optimization problem is just convex. That is, we assume that we are given (a) a Lipschitz-continuous function f defined on $\mathbb{R}^n$ (with C the domain of its Fenchel-conjugate), (b) a lower-semicontinuous and 1-strongly convex function h with *compact* domain K, which is differentiable on $\operatorname{int}(K)$ and such that for all $(x_1, x_2) \in K \times \operatorname{int}(K)$, $D(x_1, x_2) \leqslant \delta^2$, and (c) a matrix $A \in \mathbb{R}^{n \times p}$. We consider the problem,

$$\min_{x \in K} f(Ax),$$

and we let $x_* \in K$ denote any minimizer. We have the following Fenchel duality relationship:

$$\begin{aligned} \min_{x \in K} f(Ax) &= \min_{x \in K} \max_{y \in C} y^\top Ax - f^*(y) \\ &= \max_{y \in C} -\sigma(-A^\top y) - f^*(y), \end{aligned}$$

where $\sigma : \mathbb{R}^p \to \mathbb{R}$ is the support function of K defined as $\sigma(z) = \max_{x \in K} z^\top x$. We consider the mirror descent recursion [24, 4], started from $x_0 \in K$ and for which, for $t \geqslant 1$,

$$z_t \in \arg\min_{x \in K} \ \frac{1}{\rho_t} D(x, x_{t-1}) + (x - x_{t-1})^\top A^\top y_{t-1},$$

where y_{t-1} is a subgradient of f at Ax_{t-1}. Our goal is to show that the average iterate $\bar{x}_t = \frac{1}{t} \sum_{u=0}^{t-1} x_u$ and the average dual candidate $\bar{y}_t = \frac{1}{t} \sum_{u=0}^{t-1} y_u$ are such that

$$\operatorname{gap}(\bar{x}_t, \bar{y}_t) = f(A\bar{x}_t) + \sigma(-A^\top \bar{y}_t) + f^*(\bar{y}_t)$$

tends to zero at an appropriate rate. Similar results hold for certain cases of subgradient descent [23] and we show that they hold more generally.

Let $x \in K$. We have (using a similar reasoning than [4]) and using the optimality condition $(x - x_t)^\top \big[h'(x_t) - h'(x_{t-1}) + \rho_t A^\top y_{t-1}\big] \geqslant 0$ for any $x \in K$:

$$\begin{aligned} D(x, x_t) - D(x, x_{t-1}) &= -h(x_t) - h'(x_t)^\top (x - x_t) + h(x_{t-1}) + h'(x_{t-1})^\top (x - x_{t-1}) \\ &\leqslant \rho_t (x - x_{t-1})^\top A^\top y_{t-1} + h(x_{t-1}) - h(x_t) - h'(x_t)^\top (x_{t-1} - x_t) \\ &\leqslant \rho_t (x - x_{t-1})^\top A^\top y_{t-1} + \big[h'(x_{t-1}) - h'(x_t)\big]^\top (x_{t-1} - x_t) - \frac{1}{2}\|x_t - x_{t-1}\|^2 \\ &\qquad \text{using the 1-strong convexity of } h, \\ &= \rho_t (x - x_{t-1})^\top A^\top y_{t-1} + \big[\rho_t A^\top y_{t-1}\big]^\top (x_{t-1} - x_t) - \frac{1}{2}\|x_t - x_{t-1}\|^2 \\ &= \rho_t (x - x_{t-1})^\top A^\top y_{t-1} + \frac{1}{2}\|\rho_t A^\top y_{t-1}\|^2 - \frac{1}{2}\|x_t - x_{t-1} + \rho_t A^\top y_{t-1}\|^2 \\ &\leqslant \rho_t (x - x_{t-1})^\top A^\top y_{t-1} + \frac{1}{2}\|\rho_t A^\top y_{t-1}\|^2. \end{aligned}$$

This leads to

$$(x_{t-1} - x)^\top A^\top y_{t-1} \leqslant \frac{1}{\rho_t}\big[D(x, x_{t-1}) - D(x, x_t)\big] + \frac{\rho_t R^2}{2},$$

where $R^2 = \max_{y \in C} \|A^\top y\|^2$ (note the slightly different definition than in Section 4). By summing from $u = 1$ to t, we obtain

$$\sum_{u=1}^{t} (x_{u-1} - x)^\top A^\top y_{u-1} \;\leqslant\; \sum_{u=1}^{t} \frac{1}{\rho_u} \big[D(x, x_{u-1}) - D(x, x_u) \big] + \sum_{u=1}^{t} \frac{\rho_u R^2}{2}.$$

Assuming that (ρ_t) is a decreasing sequence, we get by integration by parts:

$$\begin{aligned}
\sum_{u=1}^{t} (x_{u-1} - x)^\top A^\top y_{u-1} &\leqslant \sum_{u=1}^{t-1} D(x, x_u) \Big(\frac{1}{\rho_{u+1}} - \frac{1}{\rho_u} \Big) + \frac{D(x, x_0)}{\rho_1} - \frac{D(x, x_t)}{\rho_t} + \sum_{u=1}^{t} \frac{\rho_u R^2}{2} \\
&\leqslant \sum_{u=1}^{t-1} \delta^2 \Big(\frac{1}{\rho_{u+1}} - \frac{1}{\rho_u} \Big) + \frac{\delta^2}{\rho_1} + \sum_{u=1}^{t} \frac{\rho_u R^2}{2} \\
&= \frac{\delta^2}{\rho_t} + \frac{R^2}{2} \sum_{u=1}^{t} \rho_u.
\end{aligned}$$

We may now compute an upper-bound on the gap as follows:

$$\begin{aligned}
\mathrm{gap}(\bar{x}_t, \bar{y}_t) &= f(A\bar{x}_t) + \sigma(-A^\top \bar{y}_t) + f^*(\bar{y}_t) \\
&\leqslant \frac{1}{t} \sum_{u=0}^{t-1} f(Ax_u) + \frac{1}{t} \sum_{u=0}^{t-1} f^*(y_u) + \sigma(-A^\top \bar{y}_t) \\
&= \frac{1}{t} \sum_{u=0}^{t-1} y_u^\top A x_u + \sigma(-A^\top \bar{y}_t) \\
&= \frac{1}{t} \sum_{u=0}^{t-1} y_u^\top A (x_u - x) + \sigma(-A^\top \bar{y}_t) + x^\top A^\top \bar{y}_t.
\end{aligned}$$

Using the bound above and minimizing with respect to $x \in K$, we obtain

$$\mathrm{gap}(\bar{x}_t, \bar{y}_t) \leqslant \frac{\delta^2}{t\rho_t} + \frac{R^2}{2t} \sum_{u=1}^{t} \rho_u.$$

With $\rho_t = \dfrac{\delta}{R\sqrt{t}}$, we obtain a gap less than

$$\frac{R\delta}{\sqrt{t}} + \frac{R\delta}{t} \sum_{k=1}^{t} \big[\sqrt{k} - \sqrt{k-1} \big] \leqslant \frac{2R\delta}{\sqrt{t}}.$$