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

January 2, 2013

Abstract

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 the support vector machine), 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.

In this paper, we consider two classical algorithms, namely (a) subgradient descent and its mirror descent extension [1, 2, 3], and (b) conditional gradient algorithms, sometimes referred to as Frank-Wolfe algorithms [4, 5, 6, 7, 8].

Subgradient algorithms are adapted to non-smooth situations, have a convergence rate of $O(t^{-1/2})$ in terms of objective values, after t steps. This convergence rate goes to $O(t^{-1})$ when the objective function is strongly convex [9]. Conditional-gradient algorithms are tailored to the optimization of smooth functions on a compact convex set, for which minimizing linear functions is easy (but typically, orthogonal projections would be hard, so that proximal methods [10, 11] cannot be used efficiently). They also have a convergence rate of $O(1/t)$ [6]. The main results of this paper are (a) to show that 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 [12], and (c) provide explicit convergence rates for primal and dual iterates.

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 linear operator A from $\mathbb{R}^p$ to $\mathbb{R}^n$. 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** B**-Lipschitz-continuous and finite on** $\mathbb{R}^n$, i.e., 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 included in the ball of center 0 and radius B. We denote by C the compact domain of f^*. Thus, for all $z \in \mathbb{R}^n$, $f(z) = \max_{y \in C} y^\top z - f^*(y)$.

  Note that the compactness of the domain of f^* is crucial and allows for simpler proof techniques with explicit constants (see a generalization in [12]).

- 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 [13, 14]. 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^* are useful to compute duality gaps.

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 thus well adapted to the subgradient method.

We have the following primal/dual relationships (obtained from Fenchel duality [13]):

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

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 *compact* 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.

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

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

This quantity serves as a certificate of optimality, as

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

2 Examples

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 convex set, which is μ-strongly convex.
- **Negative entropy**: $h(x) = \sum_{i=1}^n x_i \log x_i + I_S(x)$, where $S = \{x \in \mathbb{R}^n,\ x \geqslant 0,\ \sum_{i=1}^n x_i = 1\}$, which is 1-strongly convex. More generally, many barrier functions of convex sets may be used (see examples in [3, 15]).

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 $\mathbb{R}^p$ (but it is, when restricted to a compact 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. This extends to any negative exponential family log-likelihood. 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).
- **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 [16].

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 [17], or submodular function minimization cast through separable optimization problems [18].

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., [14]). 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 $x_1 = x_2$. See more details in [3]. For example, when $h(x) = \frac{\mu}{2}\|x\|^2$, we have $D(x_1, x_2) = \frac{1}{2}\|x_1 - x_2\|^2$.

Subgradient descent for square Bregman divergence. When $h(x) = \frac{\mu}{2}\|x\|^2$, the primal problem becomes:

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

The projected subgradient method starts from $x_0 \in \mathbb{R}^p$, 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 $f'(Ax_{t-1})$ 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 the minimization of

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

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

Mirror descent. We may interpret the last formulation as the minimization of

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

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

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

Thus, we now define the mirror descent recursion as follows:

$$\left\{\begin{array}{rcl} \bar{y}_{t-1} & \in & \arg\max_{y\in C} \; y^\top A x_{t-1} - f^*(y), \\ x_t & = & \arg\min_{x\in\mathbb{R}^p} \; h(x) - (1-\rho_t) x^\top h'(x_{t-1}) + \rho_t x^\top A^\top \bar{y}_{t-1}. \end{array}\right. \tag{3}$$

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. (3), we have:*

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

Proof We follow the proof of [3] and adapt it to the strongly convex case. We have:

$$\begin{array}{rl} & D(x_*, x_t) - D(x_*, x_{t-1}) \\ = & h(x_{t-1}) - h(x_t) - (x_* - x_t)^\top h'(x_t) + (x_* - x_{t-1})^\top h'(x_{t-1}) \\ = & h(x_{t-1}) - h(x_t) - (x_* - x_t)^\top \big[(1-\rho_t) h'(x_{t-1}) - \rho_t A^\top f'(A x_{t-1})\big] + (x_* - x_{t-1})^\top h'(x_{t-1}) \\ = & h(x_{t-1}) - h(x_t) - (x_{t-1} - x_t)^\top h'(x_{t-1}) + \rho_t (x_* - x_t)^\top g'_{\text{primal}}(x_{t-1}) \\ = & \big[- D(x_t, x_{t-1}) + \rho_t (x_{t-1} - x_t)^\top g'_{\text{primal}}(x_{t-1})\big] + \big[\rho_t (x_* - x_{t-1})^\top g'_{\text{primal}}(x_{t-1})\big]. \end{array} \tag{4}$$

In order to upper-bound the two terms in Eq. (4), we first consider the following bound (obtained by convexity of f):

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

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

$$\rho_t (x_* - x_{t-1})^\top g'_{\text{primal}}(x_{t-1}) \leqslant -\rho_t D(x_*, x_{t-1}) - \rho_t \big[g_{\text{primal}}(x_{t-1}) - g_{\text{primal}}(x_*)\big]. \tag{5}$$

Moreover,

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

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, $z = \rho_t(A^\top f'(Ax_{t-1}) + h(x_{t-1}))$. Since $h'(x_{t-1}) \in -A^\top K$ (because $h'(x_{t-1})$ is a convex combination of such elements), then $\|A^\top f'(Ax_{t-1}) + h(x_{t-1})\|^2 \leqslant R^2 = \max_{y_1, y_2 \in K} \|A^\top(y_1 - y_2)\|^2 = \operatorname{diam}(A^\top K)^2$.

Overall, combining Eq. (5) and $\varphi(z) \leqslant \frac{R^2\rho_t^2}{2\mu}$ into Eq. (4), 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

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

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 (a) $D(x_*, x_t) \leqslant \dfrac{R^2}{\mu(t+1)}$, i.e., the iterates converges, and (b)

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

the objective functions at an average data point converges, and (c)

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

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

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

which implies

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

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

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

that is, $h'(x_t)$ is a weighted average of subgradients.

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 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 Prop. 1 above holds.

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

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

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

4 Conditional gradient method and extensions

Conditional gradient method. Given a maximization problem of the form (i.e., where f^* is zero on its domain)

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

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):

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

It corresponds to a linearization of $-h^*(-A^\top y)$ and its maximization over the compact 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 [12], 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:

$$\left\{ \begin{array}{rcl} x_{t-1} & = & \arg\min_{x \in \mathbb{R}^p} h(x) + x^\top A^\top y_{t-1} = (h^*)'(-A^\top y_{t-1}) \\ \bar{y}_{t-1} & \in & \arg\max_{y \in C} y^\top A x_{t-1} - f^*(y) \\ y_t & = & (1-\rho_t) y_{t-1} + \rho_t \bar{y}_{t-1}. \end{array} \right. \tag{6}$$

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 [10, 11].

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:

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

which is exactly the mirror descent algorithm described in Eq. (3). 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. (3), started from $x_0 = (h^*)'(-A^\top y_0)$, is equivalent to the generalized conditional gradient recursion in Eq. (6), started from $y_0 \in C$.*

When h is not essentially smooth, then with a particular choice of subgradient, 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 compactness assumptions [12], 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:

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 mirror descent recursion of Eq. (6), we have:*

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

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

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

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 $\rho_t \in [0,1]$:

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

which is the classical equation from the conditional gradient algorithm [6, 7, 8], which we can analyze through Lemma 1 (see end of this section). ■

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 $\rho_t = \min\{\frac{\mu}{B^2}\mathrm{gap}(x_{t-1}, y_{t-1}), 1\}$ 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 mirror descent recursion of Eq. (6), we have:

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

Proof The proof is essentially the same as the previous one, with a different application of Lemma 1 (see end of this section). ■

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 $k \in \{\lfloor t/2 \rfloor, \ldots, t\}$ 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 $k \in \{\lfloor t/2 \rfloor - 1, \ldots, t\}$ 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

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

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 $t \in \{k+1, \ldots, j\}$, 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 all $v_{t-1} \geqslant \beta$ for $t \in \{k+1, \dots, j\}$, then

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

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, $u_1 \leqslant \frac{A}{2}$, 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, then

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

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), 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 of guarantees, with a gap that converges at the same rate $\frac{R^2}{\mu t}$ (see [8] for similar results for the regular conditional gradient method). These certificates may either be taken to be the pair (x_t, y_t), in which case, we have shown that after t iterations, at least one of them 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 (see, e.g., [19]).

- **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 [20].

- **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 [21, 22], which would be worth further investigating.

- **Stochastic approximation and online learning**: There are potentially other exchanges between primal/dual formulations, in particular in the stochastic setting (see, e.g., [23]).

- **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., [24]).

- **Conditional gradient algorithms for penalized problems**: Another interesting example for machine learning is more naturally described from the dual formulation: given a smooth loss term $h^*(-A^\top y)$ (this could be least-squares or logistic regression), a typically non-smooth penalization is added, often is the form of a constant times a norm, i.e., $f^*(y) = \lambda \Omega(y)$. When the proximal operator for the norm Ω is easy to compute, then the minimization of $h^*(-A^\top y) + f^*(y)$ may readily be done through proximal methods [10, 11]. However, in some situations, the only efficient operation on the norm Ω is the maximization of linear functions on the unit ball.

Conditional gradient algorithms are applicable to functions f^* with a *compact* domain and are thus adapted to constrained problems where f^* would be the indicator function of the ball $\{y \in \mathbb{R}^n,\ \Omega(y) \leqslant \nu\}$. However, the penalized problem defined above does not satisfy the compactness assumption and an extension has been recently proposed in [25]: given a (potentially loose) bound ν on an optimal solution, a line-search-based algorithm is derived that leads to a convergence rate of $O(1/t)$, with proportionally constant independent of ν. A simpler algorithm that does not exhibit this property may be obtained by considering the function defined as $f^*(y) = \lambda\Omega(y)$ for $\Omega(y) \leqslant \nu$ and $+\infty$ otherwise, which does have a compact domain, and the generalized conditional gradient algorithms described in Section 4.

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

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

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

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

[5] 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.

[6] 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.

[7] 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.

[8] M. Jaggi. Convex optimization without projection steps. Technical Report 1108.1170, Arxiv, 2011.

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

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

[11] 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.

[12] 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.

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

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

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

[16] 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.

[17] 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.

[18] F. Bach. Learning with submodular functions: A convex optimization perspective. Technical Report 1111.6453, Arxiv, 2011.

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

[20] 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.

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

[22] 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.

[23] S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. Technical Report 1207.4747, arXiv, 2012.

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

[25] Z. Harchaoui, A. Juditsky, and A. Nemirovski. Conditional gradient algorithms for machine learning. In *NIPS Workshop on Optimization*, 2012.