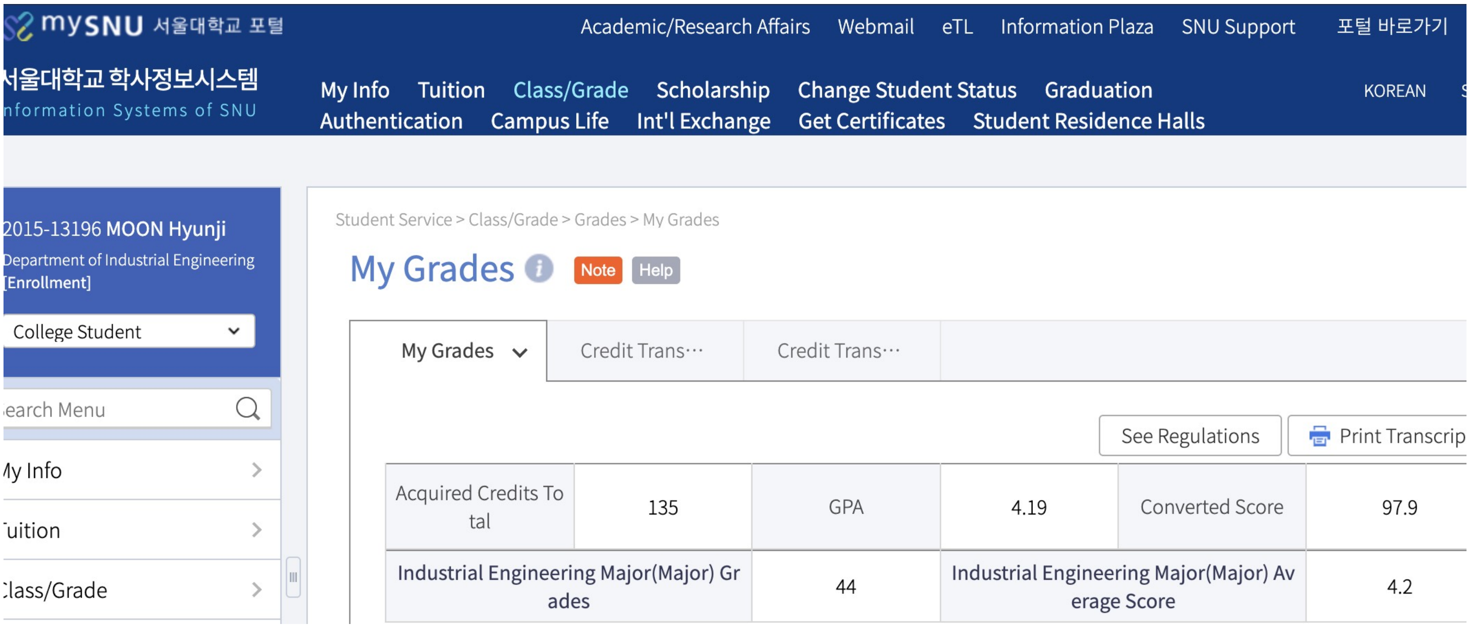

2015-13196 MOON Hyunji
Department of Industrial Engineering
[Enrollment]

My Grades

Acquired Credits Total	135	GPA	4.19	Converted Score	97.9
Industrial Engineering Major(Major) Grades	44	Industrial Engineering Major(Major) Average Score	4.2		

scree captured from webpage for overall GPA (4.2/4.3)

SIAM J. OPTIM.
Vol. 25, No. 1, pp. 115–129

DUALITY BETWEEN SUBGRADIENT AND CONDITIONAL GRADIENT METHODS*

FRANCIS BACH†

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 nonsmooth losses or problems regularized by nonsmooth 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.

Key words. subgradient, conditional gradient, nonsmooth optimization

AMS subject classifications. 90C52, 90C46, 90C06

DOI. 10.1137/130941961

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 nonsmooth parts of the objective function, when it is decomposable in a smooth and a nonsmooth part.

In this paper, we consider two classical algorithms, namely (a) subgradient descent and its mirror descent extension [4, 26, 32] and (b) conditional gradient algorithms, sometimes referred to as Frank–Wolfe algorithms [14, 15, 16, 17, 21].

Subgradient algorithms are adapted to nonsmooth 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 [24]. 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 [5, 28] cannot be used efficiently). They also have a convergence rate of $O(1/t)$ [16]. 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) to provide explicit convergence rates for primal and dual iterates.

min linear function is easy
but projection is hard

1.1. Assumptions.

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 $A \in$

*Received by the editors October 21, 2013; accepted for publication (in revised form) October 30, 2014; published electronically January 8, 2015. This work was partially supported by the European Research Council (SIERRA Project).
http://www.siam.org/journals/siopt/25-1/94196.html

†INRIA, SIERRA Project Team, Département d'Informatique de l'Ecole Normale Supérieure, Paris, France (francis.bach@ens.fr).

$\mathbb{R}^{n\times p}$. Throughout this paper, unless otherwise stated, we consider a generic norm $\|\cdot\|$ and its dual norm $\|\cdot\|_*$; it does not need to be the Euclidean norm $\|\cdot\|_2$.

We consider the following minimization problem, which we refer to as the *primal* problem:

$$\min_{x\in\mathbb{R}^p} \; h(x) + f(Ax). \tag{1.1}$$

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 $x, y \in \mathbb{R}^n$, $|f(x) - f(y)| \leqslant B\|x - y\|$. 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 as 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]). In particular, since f is B-Lipschitz-continuous, C is included in the ball of center zero and radius B. Note, however, that the quantity driving the convergence rate will be slightly different, i.e., $R = \max_{y,y'\in C} \|A^\top(y - y')\|_*$.
- *h is lower-semicontinuous and μ-strongly convex on $\mathbb{R}^p$ with respect to the norm* $\|\cdot\|$, i.e., for all $x, y \in \mathbb{R}^p$ and any subgradient $h'(x)$ of h at x, we have
  $$h(y) \geqslant h(x) + h'(x)^\top(y - x) + \frac{\mu}{2}\|x - y\|^2.$$
  This implies that h^* is defined on $\mathbb{R}^p$, differentiable, and $(1/\mu)$-smooth for the norm $\|\cdot\|_*$ [31, Lemma 2], i.e., for all $x, y \in \mathbb{R}^p$, we have
  $$h^*(y) \leqslant h^*(x) + (h^*)'(x)^\top(y - x) + \frac{1}{2\mu}\|x - y\|_*^2.$$
  Note that the domain K of h may be strictly included in $\mathbb{R}^p$.
  In most of the paper, we will assume that the function h is essentially smooth, that is, 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. This makes possible the equivalence between mirror descent and conditional gradient. However, some of our convergence results hold in more general situations, in particular when the interior of K is empty or when K is closed and h is differentiable on K (see the end of section 4 for details).

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 $\max_{y\in C} y^\top z - f^*(y) - \omega(y)$, for some strongly convex prox-function ω, then proximal methods [5, 28] applied to the dual problem converge at rate $O(1/t^2)$. If f and h are smooth, then gradient methods (accelerated [27, section 2.2] or not) have linear convergence rates.

1.2. Primal-dual relationships. We denote by $g_{\text{primal}}(x) = h(x) + f(Ax)$ the primal objective in (1.1). 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 [32].

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

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

This leads to the *dual* maximization problem:

$$\max_{y\in C} -h^*(-A^\top y) - f^*(y). \tag{1.2}$$

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 nonsmooth part $-f^*(y)$, and the problem is restricted onto a *bounded* set C. When f^* is affine (and more generally smooth) on its support, then we are exactly in the situation where conditional gradient algorithms may be used [14, 17].

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^* affine on its domain and h a squared Euclidean norm), the nonprojected subgradient method applied to the primal problem in (1.1) is equivalent to the conditional gradient applied to the dual problem in (1.2); when relaxing some of the assumptions above, this equivalence is then between mirror descent methods and generalized conditional gradient algorithms. This allows notably to transfer convergence rate analyses.

2. Examples.

The nonsmooth strongly convex optimization problem defined in (1.1) occurs in many applications in machine learning and signal processing, either because they are formulated directly in this format or their dual in (1.2) is (i.e., the original problem is the minimization of a smooth function over a compact set).

h(x) + f(AX)

2.1. Direct formulations.

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

- *Squared Euclidean norm*: $h(x) = \frac{\mu}{2}\|x\|_2^2$, which is μ-strongly convex with respect to the ℓ_2-norm $\|\cdot\|_2$. Note that the squared ℓ_1-norm is not strongly convex with respect to any norm.
- *Squared Euclidean norm with convex constraints*: $h(x) = \frac{\mu}{2}\|x\|_2^2 + I_K(x)$, with I_K the indicator function for K a closed convex set, which is μ-strongly convex with respect to the ℓ_2-norm. Note that it is not essentially smooth when $K \neq \mathbb{R}^p$.

- *Negative entropy*: $h(x) = \sum_{i=1}^{p} x_i \log x_i + I_K(x)$, where $K = \{x \in \mathbb{R}^p,\ x \geqslant 0,\ \sum_{i=1}^{p} x_i = 1\}$, which is 1-strongly convex with respect to the ℓ_1-norm [4]. More generally, many barrier functions of convex sets may be used (see examples in [4, 9], in particular for problems on matrices). This function is essentially smooth if restricted to the hyperplane $\{\sum_{i=1}^{p} x_i = 1\}$.

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., [35] 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) [5, 28].
- *Support vector machine*: $\ell_i(z_i) = \max\{1 - y_i z_i, 0\}$ with $y_i \in \{-1, 1\}$. Here f^* is affine on its domain (this is a situation where subgradient and conditional gradient methods are exactly equivalent). This extends to more general "max-margin" formulations [33, 34]: 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 affine 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 nonsmooth [11], or submodular function minimization cast through separable optimization problems [2]. Moreover, many proximal operators for nonsmooth regularizers are of this form, with $h(x) = \frac{1}{2}\|x - x_0\|_2^2$, 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 (1.2): given a smooth loss term $h^*(-A^\top y)$ (this could be least-squares or logistic regression), a typically nonsmooth 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 nondecreasing 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 [21].

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 correspond to the constraint $\Omega(y) \leqslant \omega_0$. We may also consider $\varphi(\omega) = \lambda\omega + I_{[0,\omega_0]}(\omega)$, which correspond 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 [1, 19, 37] 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 $\text{int}(K)$ to $\mathbb{R}^p$, where K is the domain of h (see, e.g., [30, 20]). Note that this imposes that K has nonempty interior; see the end of section 4 for straightforward extensions to these cases, which include the simplex. See also section 3.4 for extensions to h nonessentially smooth.

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 $K \times \text{int}(K)$ and is nonnegative. If $x_1, x_2 \in \text{int}(K)$, 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 $h(x) = \frac{\mu}{2}\|x\|_2^2$ and $\|\cdot\| = \|\cdot\|_2$, we have $D(x_1, x_2) = \frac{\mu}{2}\|x_1 - x_2\|_2^2$.

3.1. Subgradient descent for square Bregman divergence.

We first consider the common situation where $h(x) = \frac{\mu}{2}\|x\|_2^2$, with full domain $K = \mathbb{R}^p$, and $\|\cdot\| = \|\cdot\|_2$; the primal problem then becomes:

$$\min_{x\in\mathbb{R}^p} f(Ax) + \frac{\mu}{2}\|x\|_2^2.$$

The (nonprojected) 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 $f'(Ax_{t-1})$ is any subgradient of f at Ax_{t-1}. The step size is $\frac{\rho_t}{\mu}$.

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

$$(x - x_{t-1})^\top\big[A^\top f'(Ax_{t-1}) + \mu x_{t-1}\big] + \frac{\mu}{2\rho_t}\|x - x_{t-1}\|_2^2, \tag{3.1}$$

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

3.2. Mirror descent.

We may interpret the last formulation in (3.1) for the square regularizer $h(x) = \frac{\mu}{2}\|x\|_2^2$ 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 $\mathrm{int}(K)$ to $\mathbb{R}^p$)

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

This leads to the following definition of the mirror descent recursion, with $\bar{y}_{t-1}$ being any subgradient of f at Ax_{t-1}:

$$\left\{ \begin{array}{rcl} \bar{y}_{t-1} & \in & \arg\max\limits_{y \in C} \ y^\top A x_{t-1} - f^*(y), \\ x_t & = & \arg\min\limits_{x \in \mathbb{R}^p} \ h(x) - (1-\rho_t) x^\top h'(x_{t-1}) + \rho_t x^\top A^\top \bar{y}_{t-1}. \end{array} \right. \tag{3.2}$$

The following proposition proves the convergence of a specific instance of mirror descent in the strongly convex case with rate $O(1/t)$, where the strongly convex function h defining the Bregman divergence is also the direct source of strong convexity of the objective function $g_{\mathrm{primal}}(x) = h(x) + f(Ax)$. This is in contrast with earlier related works [18, 25] which consider a generic strongly convex function (independent of the choice of the Bregman divergence) and also obtain convergence rates in $O(1/t)$ with fewer assumptions for the function to minimize (no essential smoothness, possible extensions to stochastic situations) but stronger assumptions on the Bregman divergence (upper-bounded by squared distance). In this work, we chose this particular instance of mirror descent because of the equivalence with the generalized conditional gradient which we outline in section 4.

PROPOSITION 3.1 (convergence of mirror descent in the strongly convex case). *Assume that* (a) *f is Lipschitz-continuous and finite on $\mathbb{R}^n$, 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* (3.2)*, we have*

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

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

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

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

$$\begin{aligned} & D(x_*, x_t) - D(x_*, x_{t-1}) \\ &\quad = h(x_{t-1}) - h(x_t) - (x_* - x_t)^\top h'(x_t) + (x_* - x_{t-1})^\top h'(x_{t-1}) \\ &\quad = h(x_{t-1}) - h(x_t) - (x_* - x_t)^\top \big[(1-\rho_t)h'(x_{t-1}) - \rho_t A^\top f'(Ax_{t-1})\big] \\ &\qquad + (x_* - x_{t-1})^\top h'(x_{t-1}) \\ &\quad = h(x_{t-1}) - h(x_t) - (x_{t-1} - x_t)^\top h'(x_{t-1}) + \rho_t (x_* - x_t)^\top g'_{\mathrm{primal}}(x_{t-1}) \\ &\quad = \big[- D(x_t, x_{t-1}) + \rho_t (x_{t-1} - x_t)^\top g'_{\mathrm{primal}}(x_{t-1})\big] \\ &\qquad + \big[\rho_t (x_* - x_{t-1})^\top g'_{\mathrm{primal}}(x_{t-1})\big]. \end{aligned} \tag{3.3}$$

In order to upper-bound the two terms in (3.3), 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

$$\rho_t(x_*-x_{t-1})^\top g'_{\text{primal}}(x_{t-1}) \leqslant -\rho_t D(x_*,x_{t-1})-\rho_t\big[g_{\text{primal}}(x_{t-1})-g_{\text{primal}}(x_*)\big]. \tag{3.4}$$

Moreover, by the 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 with respect to $\|\cdot\|$, and its Fenchel conjugate is thus $(1/\mu)$-smooth with respect to the dual norm $\|\cdot\|_*$ [31, Lemma 2]. This implies that φ is $(1/\mu)$-smooth. Since $\varphi(0)=0$ and $\varphi'(0)=0$, we have $\varphi(z) \leqslant \frac{1}{2\mu}\|z\|_*^2$.

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 since $\rho_1 = 1$, and thus no assumption is needed regarding $h'(x_0)$), 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 (3.4) and $\varphi(z) \leqslant \frac{R^2\rho_t^2}{2\mu}$ into (3.3), 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 $\rho_t = \frac{2}{t+1}$, 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 $D(x_*,x_t) \leqslant \frac{R^2}{\mu(t+1)}$, i.e., the iterates converge. Moreover, using the convexity of g,

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

i.e., the objective functions at an averaged iterate converge, 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. ▯

Dealing with empty interiors. Throughout section 3.2, we have assumed that h is essentially smooth, which implies that its domain K has a nonempty interior. Unfortunately, several cases of interest, such as the simplex, do have an empty interior. However, Proposition 3.1 still applies when h is assumed relatively smooth when restricted to the affine hull of its domain. This is a direct consequence of the affine invariance of the recursion in (3.2).

3.3. Averaging. Note that with the step size $\rho_t = \frac{2}{t+1}$, 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.,

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

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

For $\rho_t = 1/t$, when using the same technique, 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 [22]). We would then have $h'(x_t) = \frac{1}{t}\sum_{u=1}^{t}\big[-A^\top f'(Ax_{u-1})\big]$, and this is exactly a form dual averaging method [29], which also comes with primal-dual guarantees.

3.4. Generalization to *h* nonessentially smooth. The previous result does not require h to be essentially smooth, i.e., it may be applied to $h(x) = \frac{\mu}{2}\|x\|_2^2 + I_K(x)$ and $\|\cdot\| = \|\cdot\|_2$, where K is a closed convex set strictly included in $\mathbb{R}^p$. In the mirror descent recursion,

$$\begin{cases} \bar{y}_{t-1} & \in & \arg\max_{y\in C}\ y^\top Ax_{t-1} - f^*(y), \\ x_t & = & \arg\min_{x\in\mathbb{R}^p}\ h(x) - (1-\rho_t)x^\top h'(x_{t-1}) + \rho_t x^\top A^\top \bar{y}_{t-1}, \end{cases}$$

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 Proposition 3.1 above holds.

Note that when $h(x) = \frac{\mu}{2}\|x\|_2^2 + I_K(x)$, the algorithm above is *not* equivalent to classical projected gradient descent. Indeed, the classical algorithm has the iteration

$$\begin{aligned} x_t &= \Pi_K\left(x_{t-1} - \frac{1}{\mu}\rho_t\big[\mu x_{t-1} + A^\top f'(Ax_{t-1})\big]\right) \\ &= \Pi_K\left((1-\rho_t)x_{t-1} + \rho_t\left[-\frac{1}{\mu}A^\top f'(Ax_{t-1})\right]\right) \end{aligned}$$

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 and Lipschitz-continuous, then we may modify the proof to obtain a similar result. Indeed, the bound of Proposition 3.1 still holds because the optimality condition $h'(x_t) = h'(x_{t-1}) - \rho_t[A^\top f'(Ax_{t-1}) + h'(x_{t-1})]$ may now be replaced by $(x - x_t)^\top(h'(x_t) - h'(x_{t-1}) + \rho_t[A^\top f'(Ax_{t-1}) + h'(x_{t-1})]) \geqslant 0$ for all $x \in K$, which also allows us to get to (3.3) in the proof of Proposition 3.1. The only other place where the essential smoothness of h is needed is when we use the fact that $h'(x_{t-1}) \in -A^\top C$ to obtain an upper bound on $\|A^\top f'(Ax_{t-1}) + h(x_{t-1})\|_*^2$ which is equal to the squared diameter of $A^\top C$. We now need to replace this squared diameter by $\sup_{x\in K}\|h'(x) + A^\top f'(Ax)\|_*^2$, which is finite because both h and f are now assumed Lipschitz-continuous. Note that we could use the tools from [23, Appendix A] to obtain a sharper constant.

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 affine in its domain, and then we present its generalization.

4.1. Conditional gradient method.

When f^* is affine on its domain, then by a simple change of variable, we may assume without loss of generality that f^* is zero on its domain (i.e., an indicator function).[1]

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 and that we are maximizing the first-order Taylor expansion to obtain a candidate $\bar{y}_{t-1}$ toward which we make a small step):

$$\begin{aligned} x_{t-1} &= \arg\min_{x\in\mathbb{R}^p} h(x) + x^\top A^\top y_{t-1}, \\ \bar{y}_{t-1} &\in \arg\max_{y\in C} y^\top A x_{t-1}, \\ y_t &= (1-\rho_t)y_{t-1} + \rho_t\bar{y}_{t-1}. \end{aligned}$$

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 or by (approximate) line search.

[1] Indeed, if $f^*(y) = a^\top y + b$ on its domain, then g defined by $g(x) = f(x+a) + b$ has a Fenchel conjugate which is zero on its domain. This does change the problem but not our convergence results, since our algorithms and results are invariant by translation in x.

4.2. 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:

$$(4.1) \qquad \begin{cases} x_{t-1} &= \arg\min_{x \in \mathbb{R}^p} h(x) + x^\top A^\top y_{t-1} = (h^*)'(-A^\top y_{t-1}) \\ \bar{y}_{t-1} &\in \arg\max_{y \in C} y^\top A x_{t-1} - f^*(y), \\ y_t &= (1-\rho_t) y_{t-1} + \rho_t \bar{y}_{t-1}. \end{cases}$$

Note that given our assumptions, the algorithm is always well-defined (in particular the second step with the maximization on C). This is obvious when C is closed, but the closedness of C is not explicitly among our assumptions. When f^* is affine on its domain, then it is implied by our assumption that f has full domain; in general, we need not that C is closed but that the step $\bar{y}_{t-1} \in \arg\max_{y \in C} y^\top A x_{t-1} - f^*(y)$ is always defined. The step is equivalent to $\bar{y}_{t-1} \in \partial f(Ax_{t-1})$ the subdifferential of f at Ax_{t-1}, which we have assumed to be nonempty.

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 toward the maximizer. Note the similarity (and dissimilarity) with proximal methods which would typically add a proximal term proportional to $\|y - y_{t-1}\|_2^2$, leading to faster convergences, but with the extra requirement of solving the proximal step [5, 28].

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

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

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:

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

which is exactly the mirror descent algorithm described in (3.2). In order to have matching algorithms, we thus simply need the initial conditions to match. This leads to the following proposition.

Proposition 4.1 (equivalence between mirror descent and generalized conditional gradient). *Assume that* (a) *f is Lipschitz-continuous and finite on $\mathbb{R}^n$, with C the domain of f^*, and* (b) *h is μ-strongly convex and essentially smooth. The mirror descent recursion in* (3.2)*, started from $x_0 = (h^*)'(-A^\top y_0)$, is equivalent to the generalized conditional gradient recursion in* (4.1)*, started from $y_0 \in C$.*

When h is not essentially smooth, then with a particular choice of subgradient (see section 3.4), the two algorithms are also equivalent. We now provide convergence proofs for the two versions (with adaptive and nonadaptive 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 the generalized conditional gradient with no line search (the proof of dual convergence uses standard arguments from [14, 16], while the convergence of gaps is due to [21] for the regular conditional gradient).

Proposition 4.2 (convergence of extended conditional gradient—no line search). *Assume that* (a) *f is Lipschitz-continuous and finite on $\mathbb{R}^n$, 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* (4.1)*, we have*

$$g_{\mathrm{dual}}(y_*) - g_{\mathrm{dual}}(y_t) \leqslant \frac{2R^2}{\mu(t+1)},$$

$$\min_{u \in \{0,\ldots,t-1\}} \mathrm{gap}(x_t, y_t) \leqslant \frac{8R^2}{\mu(t+1)}.$$

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

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

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

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

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

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

which is the classical equation from the conditional gradient algorithm [15, 16, 21], which we can analyze through Lemma 4.4 (see the 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 4.3 (convergence of extended conditional gradient—with line search). *Assume that* (a) *f is Lipschitz-continuous and finite on $\mathbb{R}^n$, with C the domain of f^*, and* (b) *h is μ-strongly convex. Consider $\rho_t = \min\{\frac{\mu}{R^2}\mathrm{gap}(x_{t-1}, y_{t-1}), 1\}$ 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 generalized conditional gradient recursion of* (4.1)*, we have*

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

Proof. The proof is essentially the same as the one from the previous proposition, with a different application of Lemma 4.4 (see below). □

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

Lemma 4.4. *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 (nonadaptive sequence ρ_t), we have $\rho_1 = 1$ 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

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

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

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

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 Proposition 4.1 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 [21, 22] 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. Note the similarity with the nonstrongly convex situation where the similar uniform averages for mirror descent also have primal-dual guarantees [13].

- Line-search for mirror descent: Proposition 4.3 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 nonstrongly 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 [23].
- 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 [3, 36], 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., [18, 22, 25]).
- 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, 22] and [2, Chapter 7]).

Acknowledgments. 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. The comments of the reviewers were greatly appreciated.

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*, Found. Trends Machine Learning, 6 (2013), pp. 145–373.

[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*, Oper. Res. Lett., 31 (2003), pp. 167–175.

[5] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.

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

[7] D. P. Bertsekas and H. Yu, *A unifying polyhedral approximation framework for convex optimization*, SIAM J. Optim., 21 (2011), pp. 333–360.

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

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

[10] K. Bredies and D. A. Lorenz, *Iterated hard shrinkage for minimization problems with sparsity constraints*, SIAM J. Sci. Comput., 30 (2008), pp. 657–683.

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

[12] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, *The convex geometry of linear inverse problems*, Found. Comput. Math., 12 (2012), pp. 805–849.

[13] B. Cox, A. Juditsky, and A. Nemirovski, *Dual subgradient algorithms for large-scale nonsmooth learning problems*, Math. Program., 148 (2013), pp. 143–180.

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

[15] J. C. Dunn, *Convergence rates for conditional gradient sequences generated by implicit step length rules*, SIAM J. Control Optim., 18 (1980), pp. 473–487.

[16] J. C. Dunn and S. Harshbarger, *Conditional gradient algorithms with open loop step size rules*, J. Math. Anal. Appl., 62 (1978), pp. 432–444.

[17] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1956), pp. 95–110.

[18] S. Ghadimi and G. Lan, *Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization* I*: A generic algorithmic framework*, SIAM J. Optim., 22 (2012), pp. 1469–1492.

[19] Z. Harchaoui, A. Juditsky, and A. Nemirovski, *Conditional Gradient Algorithms for Norm-Regularized Smooth Convex Optimization*, arXiv:1302.2325, 2013.

[20] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms: Part* 1*: Fundamentals*, vol. 1, Springer, New York, 1996.

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

[22] 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.

[23] 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*, arXiv:1212.2002, 2012.

[24] A. Nedic and D. Bertsekas, *Convergence rate of incremental subgradient algorithms*, in Stochastic Optimization: Algorithms and Applications, Kluwer, Norwell, MA, 2000, pp. 263–304.

[25] A. Nedic and S. Lee, *On stochastic subgradient mirror-descent algorithm with weighted averaging*, SIAM J. Optim., 24 (2014), pp. 84–107.

[26] A. S. Nemirovski and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, John Wiley, New York, 1983.

[27] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer Academic Publishers, Norwell, MA, 2004.

[28] Y. Nesterov, *Gradient Methods for Minimizing Composite Objective Function*, Tech. report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, 2007.

[29] Y. Nesterov, *Primal-dual subgradient methods for convex problems*, Math. Program., 120 (2009), pp. 221–259.

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

[31] S. Shalev-Shwartz and Y. Singer, *Convex repeated games and Fenchel duality*, in Proceedings of Advances in Neural Information Processing Systems (NIPS), 2006.

[32] N. Z. Shor, K. C. Kiwiel, and A. Ruszczynski, *Minimization Methods for Non-differentiable Functions*, Springer-Verlag, Berlin, 1985.

[33] 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.

[34] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, *Large margin methods for structured and interdependent output variables*, J. Mach. Learn. Res., 6 (2006), pp. 1453–1484.

[35] M. J. Wainwright and M. I. Jordan, *Graphical models, exponential families, and variational inference*, Found. Trends Machine Learning, 1 (2008), pp. 1–305.

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

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