Frank-Wolfe Algorithms for Saddle Point Problems

Gauthier Gidel
INRIA - SIERRA Project-team
École normale supérieure, Paris

Tony Jebara
Department of Computer Science
Columbia University, NYC

Simon Lacoste-Julien
Department of CS & OR (DIRO)
Université de Montréal, Montréal

Abstract

We extend the Frank-Wolfe (FW) optimization algorithm to solve constrained smooth convex-concave saddle point (SP) problems. Remarkably, the method only requires access to linear minimization oracles. Leveraging recent advances in FW optimization, we provide the first proof of convergence of a FW-type saddle point solver over polytopes, thereby partially answering a 30 year-old conjecture. We also survey other convergence results and highlight gaps in the theoretical underpinnings of FW-style algorithms. Motivating applications without known efficient alternatives are explored through structured prediction with combinatorial penalties as well as games over matching polytopes involving an exponential number of constraints.

1 Introduction

The Frank-Wolfe (FW) optimization algorithm (Frank and Wolfe, 1956), also known as the conditional gradient method (Demyanov and Rubinov, 1970), is a first-order method for smooth constrained optimization over a compact set. It has recently enjoyed a surge in popularity thanks to its ability to cheaply exploit the structured constraint sets appearing in machine learning applications (Jaggi, 2013; Lacoste-Julien and Jaggi, 2015). A known forte of FW is that it only requires access to a *linear minimization oracle* (LMO) over the constraint set, i.e., the ability to minimize linear functions over the set, in contrast to projected gradient methods which require the minimization of *quadratic* functions or other nonlinear functions. In this paper, we extend the applicability of the FW algorithm to solve the following convex-concave saddle point problems:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}), \tag{1}$$

$$\text{with only access to LMO}(\mathbf{r}) \in \arg\min_{\boldsymbol{s}\in\mathcal{X}\times\mathcal{Y}} \langle \boldsymbol{s}, \mathbf{r}\rangle\,,$$

where $\mathcal{L}$ is a smooth (with L-Lipschitz continuous gradient) *convex-concave function*, that is, $\mathcal{L}(\cdot, \boldsymbol{y})$ is convex for all $\boldsymbol{y} \in \mathcal{Y}$ and $\mathcal{L}(\boldsymbol{x}, \cdot)$ is concave for all $\boldsymbol{x} \in \mathcal{X}$. We also assume that $\mathcal{X} \times \mathcal{Y}$ is a convex compact set such that its LMO is cheap to compute. A *saddle point solution* to (1) is a pair $(\boldsymbol{x}^*, \boldsymbol{y}^*) \in \mathcal{X} \times \mathcal{Y}$ (Hiriart-Urruty and Lemaréchal, 2013, VII.4) such that: $\forall \boldsymbol{x} \in \mathcal{X}$, $\forall \boldsymbol{y} \in \mathcal{Y}$

$$\mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}) \leq \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}^*) \leq \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^*). \tag{2}$$

Examples of saddle point problems. Taskar et al. (2006) cast the maximum-margin estimation of structured output models as a bilinear saddle point problem $\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top M \boldsymbol{y}$, where $\mathcal{X}$ is the regularized set of parameters and $\mathcal{Y}$ is an encoding of the set of possible structured outputs. They considered settings where projection on $\mathcal{X}$ and $\mathcal{Y}$ were efficient but one can imagine many situations where only LMO's are efficient. For example, we could use a structured sparsity inducing norm (Martins et al., 2011) for the parameter $\boldsymbol{x}$, such as the overlapping group lasso for which the projection is expensive (Bach et al., 2012), while $\mathcal{Y}$ could be a combinatorial object such as a the ground state of a planar Ising model (without external field) which admits an efficient oracle (Barahona, 1982) but has potentially intractable projection.

Similarly, two-player games (Von Neumann and Morgenstern, 1944) can often be solved as bilinear minimax problems. When a strategy space involves a polynomial number of constraints, the equilibria of such games can be solved efficiently (Koller et al., 1994). However, in situations such as the Colonel Blotto game or the Matching Duel (Ahmadinejad et al., 2016) the strategy space is intractably large and defined by an exponential number of linear constraints. Fortunately, some linear minimization oracles such as the blossom algorithm (Edmonds, 1965) can efficiently optimize over matching polytopes despite an exponential number of linear constraints.

Robust learning is also often cast as a saddle point minimax problem (Kim et al., 2005). Once again, a FW implementation could leverage fast linear oracles

while projection methods would be plagued by slower or intractable sub-problems. For instance, if the LMO is max-flow, it could have almost linear runtime while the corresponding projection would require cubic runtime quadratic programming (Kelner et al., 2014).

Related work. The standard approaches to solve smooth constrained saddle point problems are projection-type methods (surveyed in Xiu and Zhang (2003)), with in particular variations of Korpelevich's extragradient method (Korpelevich, 1976), such as (Nesterov, 2007) which was used to solve the structured prediction problem (Taskar et al., 2006) mentioned above. There is surprisingly little work on FW-type methods for saddle point problems, although they were briefly considered for the more general *variational inequality* problem (VIP):

$$\text{find} \quad \boldsymbol{z}^* \in \mathcal{Z} \quad \text{s.t.} \quad \langle \boldsymbol{r}(\boldsymbol{z}^*), \boldsymbol{z} - \boldsymbol{z}^* \rangle \geq 0, \quad \forall \boldsymbol{z} \in \mathcal{Z}, \quad (3)$$

where $\boldsymbol{r}$ is a Lipschitz mapping from $\mathbb{R}^p$ to itself and $\mathcal{Z} \subseteq \mathbb{R}^p$. By using $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\boldsymbol{r}(\boldsymbol{z}) = (\nabla_x \mathcal{L}(\boldsymbol{z}), -\nabla_y \mathcal{L}(\boldsymbol{z}))$, the variational inequality problem (3) reduces to the equivalent optimality conditions for the saddle point problem (1). Hammond (1984) showed that a FW algorithm with a step size of $O(1/t)$ converges for the VIP (3) when the set $\mathcal{Z}$ is strongly convex, while FW with a generalized line-search on a saddle point problem is sometimes non-convergent when $\mathcal{Z}$ is a polytope (see also (Patriksson, 1999, § 3.1.1)). She conjectured though that using a step size of $O(1/t)$ was also convergent when $\mathcal{Z}$ is a polytope – a problem left open up to this point. More recently, Juditsky and Nemirovski (2016) (see also Cox et al. (2015)) proposed a method to transform a VIP on $\mathcal{Z}$ where one has only access to a LMO, to a "dual" VIP on which they can use a projection-type method. Lan (2013) proposes to solve the saddle point problem (1) by running FW on $\mathcal{X}$ on the *smoothed* version of the problem $\max_{\boldsymbol{y} \in \mathcal{Y}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$, thus requiring a projection oracle on $\mathcal{Y}$. In contrast, in this paper we study simple approaches that do not require any transformations of the problem (1) nor any projection oracle on $\mathcal{X}$ or $\mathcal{Y}$.

Contributions. In § 2, we extend several variants of the FW algorithm to solve the saddle point problem (1) that we think could be of interest to the machine learning community. In § 3, we give a first proof of (geometric) convergence for these methods over polytope domains under the assumptions of sufficient strong convex-concavity of $\mathcal{L}$, giving a partial answer to the conjecture from Hammond (1984). In § 4, we extend and refine the previous convergence results when $\mathcal{X}$ and $\mathcal{Y}$ are strongly convex sets and the gradient of $\mathcal{L}$ is non-zero over $\mathcal{X} \times \mathcal{Y}$, while we survey the pure bilinear case in § 5. We finally present illustrative experiments for our theory in § 6, noticing that the convergence theory is still incomplete for these methods.

2 Saddle point Frank-Wolfe (SP-FW)

The algorithms. This article will explore three SP extensions of the classical *Frank-Wolfe* algorithm (Algorithm 1) which are summarized in Algorithm 2, 3 and 4. In the following, the point computed by these algorithms after t steps will be noted $\boldsymbol{z}^{(t)} = (\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})$. We first obtain the *saddle point FW* (SP-FW) algorithm (Algorithm 2) by simultaneously doing a FW update on both convex functions $\mathcal{L}(\cdot, \boldsymbol{y}^{(t)})$ and $-\mathcal{L}(\boldsymbol{x}^{(t)}, \cdot)$ with a properly chosen step size. Hence the point $\boldsymbol{z}^{(t)}$ has a sparse representation as a convex combination of the points previously given by the oracle. This set of points is called the *active set*. If we assume that $\mathcal{X}$ and $\mathcal{Y}$ are the convex hulls of two finite sets of points $\mathcal{A}$ and $\mathcal{B}$, we can also extend the *away-step Frank-Wolfe* (AFW) algorithm (Lacoste-Julien and Jaggi, 2015) to saddle point problems. As for AFW, this new algorithm is able to remove mass from "bad" atoms in the active set, (see L9 of Algorithm 3 and in Appendix A) to avoid the zig-zagging problem that slows down standard FW (Lacoste-Julien and Jaggi, 2015). Because of the special product structure of the domain, we consider more away directions than proposed in (Lacoste-Julien and Jaggi, 2015) for AFW. Namely, for every corner $\boldsymbol{v} = (\boldsymbol{v}_x, \boldsymbol{v}_y)$ and $\boldsymbol{v}' = (\boldsymbol{v}'_x, \boldsymbol{v}'_y)$ already picked, $\boldsymbol{x} - \boldsymbol{v}_x$ is a feasible directions in $\mathcal{X}$ and $\boldsymbol{y} - \boldsymbol{v}'_y$ is a feasible direction in $\mathcal{Y}$. Thus the combination $(\boldsymbol{x} - \boldsymbol{v}_x, \boldsymbol{y} - \boldsymbol{v}'_y)$ is a feasible direction even if the particular corners $\boldsymbol{v}_x$ and $\boldsymbol{v}'_y$ have never been picked together. We thus maintain the iterates on $\mathcal{X}$ and $\mathcal{Y}$ as independent convex combination of their respective active sets of corners (Line 13 of Algorithm 3), i.e.,

$$\boldsymbol{x}^{(t)} = \sum_{\boldsymbol{v}_x \in \mathcal{S}_x^{(t)}} \alpha_{\boldsymbol{v}_x} \boldsymbol{v}_x \quad \text{and} \quad \boldsymbol{y}^{(t)} = \sum_{\boldsymbol{v}_y \in \mathcal{S}_y^{(t)}} \alpha_{\boldsymbol{v}_y} \boldsymbol{v}_y. \quad (4)$$

Finally, a straightforward saddle point generalization for the *pairwise Frank-Wolfe* (PFW) algorithm is given in Algorithm 4. The proposed algorithms preserve several nice properties of previous FW methods (in addition to only requiring LMO's): simplicity of implementation, affine invariance (Jaggi, 2013), gap certificates computed for free, sparse representation of the iterates and the possibility to have adaptive step sizes using the gap computation. We next analyze the convergence of these algorithms.

The suboptimality error and the gap. To establish convergence, we first define several quantities of interest. In classical convex optimization, the suboptimality error h_t is well defined as $h_t := f(\boldsymbol{x}^{(t)}) - \min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$. This quantity is clearly non-negative and

Algorithm 1 Frank-Wolfe algorithm

Let $\boldsymbol{x}^{(0)} \in \mathcal{X}$
for $t = 0 \dots T$ **do**
  Compute $\boldsymbol{r}^{(t)} = \nabla f(\boldsymbol{x}^{(t)})$
  Compute $\boldsymbol{s}^{(t)} := \underset{\boldsymbol{s} \in \mathcal{X}}{\operatorname{argmin}} \left\langle \boldsymbol{s}, \boldsymbol{r}^{(t)} \right\rangle$
  Compute $g_t := \left\langle \boldsymbol{x}^{(t)} - \boldsymbol{s}^{(t)}, \boldsymbol{r}^{(t)} \right\rangle$
  if $g_t \leq \epsilon$ **then return** $\boldsymbol{x}^{(t)}$
  Let $\gamma = \frac{2}{2+t}$ (or do line-search)
  Update $\boldsymbol{x}^{(t+1)} := (1-\gamma)\boldsymbol{x}^{(t)} + \gamma \boldsymbol{s}^{(t)}$
end for

Algorithm 2 Saddle point Frank-Wolfe algorithm

Let $\boldsymbol{z}^{(0)} = (\boldsymbol{x}^{(0)}, \boldsymbol{y}^{(0)}) \in \mathcal{X} \times \mathcal{Y}$
for $t = 0 \dots T$ **do**
  Compute $\boldsymbol{r}^{(t)} := \begin{pmatrix} \nabla_x \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \\ -\nabla_y \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \end{pmatrix}$
  Compute $\boldsymbol{s}^{(t)} := \underset{\boldsymbol{z} \in \mathcal{X} \times \mathcal{Y}}{\operatorname{argmin}} \left\langle \boldsymbol{z}, \boldsymbol{r}^{(t)} \right\rangle$
  Compute $g_t := \left\langle \boldsymbol{z}^{(t)} - \boldsymbol{s}^{(t)}, \boldsymbol{r}^{(t)} \right\rangle$
  if $g_t \leq \epsilon$ **then return** $\boldsymbol{z}^{(t)}$
  Let $\gamma = \min\left(1, \frac{\nu}{2C} g_t\right)$ **or** $\gamma = \frac{2}{2+t}$
  Update $\boldsymbol{z}^{(t+1)} := (1-\gamma)\boldsymbol{z}^{(t)} + \gamma \boldsymbol{s}^{(t)}$
end for

Algorithm 3 Saddle point away-step Frank-Wolfe algorithm: **SP-AFW**$(\boldsymbol{z}^{(0)}, \mathcal{A} \times \mathcal{B}, \epsilon)$

Let $\boldsymbol{z}^{(0)} = (\boldsymbol{x}^{(0)}, \boldsymbol{y}^{(0)}) \in \mathcal{A} \times \mathcal{B}$, $\mathcal{S}_x^{(0)} := \{\boldsymbol{x}^{(0)}\}$ and $\mathcal{S}_y^{(0)} := \{\boldsymbol{y}^{(0)}\}$
for $t = 0 \dots T$ **do**
  Let $\boldsymbol{s}^{(t)} := \mathrm{LMO}_{\mathcal{A}\times\mathcal{B}}\left(\boldsymbol{r}^{(t)}\right)$ and $\boldsymbol{d}_{\mathrm{FW}}^{(t)} := \boldsymbol{s}^{(t)} - \boldsymbol{z}^{(t)}$ *($\boldsymbol{r}^{(t)}$ as defined in L3 in Algorithm 2)*
  Let $\boldsymbol{v}^{(t)} \in \underset{\boldsymbol{v} \in \mathcal{S}_x^{(t)} \times \mathcal{S}_y^{(t)}}{\arg\max} \left\langle \boldsymbol{r}^{(t)}, \boldsymbol{v} \right\rangle$ and $\boldsymbol{d}_{\mathrm{A}}^{(t)} := \boldsymbol{z}^{(t)} - \boldsymbol{v}^{(t)}$ *(the away direction)*
  if $g_t^{\mathrm{FW}} := \left\langle -\boldsymbol{r}^{(t)}, \boldsymbol{d}_{\mathrm{FW}}^{(t)} \right\rangle \leq \epsilon$ **then return** $\boldsymbol{z}^{(t)}$ *(FW gap is small enough, so return)*
  if $\left\langle -\boldsymbol{r}^{(t)}, \boldsymbol{d}_{\mathrm{FW}}^{(t)} \right\rangle \geq \left\langle -\boldsymbol{r}^{(t)}, \boldsymbol{d}_{\mathrm{A}}^{(t)} \right\rangle$ **then**
    $\boldsymbol{d}^{(t)} := \boldsymbol{d}_{\mathrm{FW}}^{(t)}$, and $\gamma_{\max} := 1$ *(choose the FW direction)*
  else
    $\boldsymbol{d}^{(t)} := \boldsymbol{d}_{\mathrm{A}}^{(t)}$, and $\gamma_{\max} := \min\left\{ \frac{\alpha_{\boldsymbol{v}_x^{(t)}}}{1-\alpha_{\boldsymbol{v}_x^{(t)}}}, \frac{\alpha_{\boldsymbol{v}_y^{(t)}}}{1-\alpha_{\boldsymbol{v}_y^{(t)}}} \right\}$ *(maximum feasible step size; a* drop step *is when $\gamma_t = \gamma_{\max}$)*
  end if
  Let $g_t^{\mathrm{PFW}} = \left\langle -\boldsymbol{r}^{(t)}, \boldsymbol{d}_{\mathrm{FW}}^{(t)} + \boldsymbol{d}_{\mathrm{A}}^{(t)} \right\rangle$ **and** $\gamma_t = \min\left\{ \gamma_{\max}, \frac{\nu^{\mathrm{PFW}}}{2C} g_t^{\mathrm{PFW}} \right\}$ *(ν^{PFW} and C set as in Thm. 1)*
  Update $\boldsymbol{z}^{(t+1)} := \boldsymbol{z}^{(t)} + \gamma_t \boldsymbol{d}^{(t)}$ *(and accordingly for the weights $\boldsymbol{\alpha}^{(t+1)}$, see Lacoste-Julien and Jaggi (2015))*
  Update $\mathcal{S}_x^{(t+1)} := \{\boldsymbol{v}_x \in \mathcal{A} \text{ s.t. } \alpha_{\boldsymbol{v}_x}^{(t+1)} > 0\}$ and $\mathcal{S}_y^{(t+1)} := \{\boldsymbol{v}_y \in \mathcal{B} \text{ s.t. } \alpha_{\boldsymbol{v}_y}^{(t+1)} > 0\}$
end for

Algorithm 4 Saddle point pairwise Frank-Wolfe algorithm: **SP-PFW**$(\boldsymbol{z}^{(0)}, \mathcal{A} \times \mathcal{B}, \epsilon)$

In Alg. 3, replace L6 to 10 by: $\boldsymbol{d}^{(t)} := \boldsymbol{d}_{\mathrm{PFW}}^{(t)} := \boldsymbol{s}^{(t)} - \boldsymbol{v}^{(t)}$, and $\gamma_{\max} := \min\left\{ \alpha_{\boldsymbol{v}_x^{(t)}}, \alpha_{\boldsymbol{v}_x^{(t)}} \right\}$.

proving that h_t goes to 0 is enough to establish convergence. Unfortunately, in the saddle point setting the quantity $\mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) - \mathcal{L}^*$ is no longer non-negative and can be equal to zero for an infinite number of points $(\boldsymbol{x}, \boldsymbol{y})$ while $(\boldsymbol{x}, \boldsymbol{y}) \notin (\mathcal{X}^*, \mathcal{Y}^*)$. For instance, if $\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x} \cdot \boldsymbol{y}$ with $\mathcal{X} = \mathcal{Y} = [-1, 1]$, then $\mathcal{L}^* = 0$ and $(\mathcal{X}^*, \mathcal{Y}^*) = \{(0, 0)\}$. But for all $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}$, $\boldsymbol{x} \cdot 0 = 0 \cdot \boldsymbol{y} = \mathcal{L}^*$. The saddle point literature thus considers a non-negative gap function (also known as a merit function (Larsson and Patriksson, 1994; Zhu and Marcotte, 1998) and (Patriksson, 1999, Sec 4.4.1)) which is zero only for optimal points, in order to quantify progress towards the saddle point. We can define the following *suboptimality error* h_t for our saddle point problem:

$$\begin{aligned} h_t &:= \mathcal{L}(\boldsymbol{x}^{(t)}, \widehat{\boldsymbol{y}}^{(t)}) - \mathcal{L}(\widehat{\boldsymbol{x}}^{(t)}, \boldsymbol{y}^{(t)}), \\ \text{where} \quad \widehat{\boldsymbol{x}}^{(t)} &:= \arg\min_{\boldsymbol{x} \in \mathcal{X}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^{(t)}), \\ \text{and} \quad \widehat{\boldsymbol{y}}^{(t)} &:= \arg\max_{\boldsymbol{y} \in \mathcal{Y}} \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}). \end{aligned} \tag{5}$$

This is an example of *primal-dual* gap function by noticing that

$$\begin{aligned} h_t &= \mathcal{L}(\widehat{\boldsymbol{y}}^{(t)}, \boldsymbol{x}^{(t)}) - \mathcal{L}^* + \mathcal{L}^* - \mathcal{L}(\boldsymbol{y}^{(t)}, \widehat{\boldsymbol{x}}^{(t)}) \\ &= p(\boldsymbol{x}^{(t)}) - p(\boldsymbol{x}^*) + g(\boldsymbol{y}^*) - g(\boldsymbol{y}^{(t)}), \end{aligned} \tag{6}$$

where $p(\boldsymbol{x}) := \max_{\boldsymbol{y} \in \mathcal{Y}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$ is the convex primal function and $g(\boldsymbol{y}) := \min_{\boldsymbol{x} \in \mathcal{X}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$ is the concave dual function. By convex-concavity, h_t can be upper-bounded by the following FW linearization gap (Jaggi,

2011, 2013; Larsson and Patriksson, 1994; Zhu and Marcotte, 1998):

$$g_t^{\mathrm{FW}} := \max_{\boldsymbol{s}_x \in \mathcal{X}} \left\{ \left\langle \boldsymbol{x}^{(t)} - \boldsymbol{s}_x, \nabla_x \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \right\rangle \right\} := g_t^{(x)} + \max_{\boldsymbol{s}_y \in \mathcal{Y}} \left\{ \left\langle \boldsymbol{y}^{(t)} - \boldsymbol{s}_y, -\nabla_y \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \right\rangle \right\} := g_t^{(y)}. \tag{7}$$

This gap is easy to compute and gives a stopping criterion since $g_t^{\mathrm{FW}} \geq h_t$.

Compensation phenomenon and difficulty for SP. Even when equipped with a suboptimality error and a gap function (as in the convex case), we still cannot apply the standard FW convergence analysis. The usual FW proof sketch uses the fact that the gradient of f is Lipschitz continuous to get

$$h_{t+1} \leq h_t - \gamma_t g_t^{\mathrm{FW}} + \gamma_t^2 \frac{L\|\boldsymbol{d}^{(t)}\|^2}{2} \tag{8}$$

which then provides a rate of convergence. Roughly, since $g_t \geq h_t$ by convexity, if γ_t is small enough then (h_t) will decrease and converge. For simplicity, in the main paper, $\|\cdot\|$ will refer to the ℓ_2 norm of $\mathbb{R}^d$. The partial Lipschitz constants and the diameters of the sets are defined with respect to this norm (see (40) in Appendix B.1 for more general norms).

Using the L-Lipschitz continuity of $\mathcal{L}$ and letting $\mathcal{L}_t := \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})$ as a shorthand, we get

$$\mathcal{L}_{t+1} \leq \mathcal{L}_t + \gamma_t \left\langle \boldsymbol{d}_x^{(t)}, \nabla_x \mathcal{L}_t \right\rangle + \gamma_t \left\langle \boldsymbol{d}_y^{(t)}, \nabla_y \mathcal{L}_t \right\rangle + \gamma_t^2 \frac{L\|\boldsymbol{d}^{(t)}\|^2}{2} \tag{9}$$

where $\boldsymbol{d}_x^{(t)} = \boldsymbol{s}_x^{(t)} - \boldsymbol{x}^{(t)}$ and $\boldsymbol{d}_y^{(t)} = \boldsymbol{s}_y^{(t)} - \boldsymbol{y}^{(t)}$. Then

$$\mathcal{L}_{t+1} - \mathcal{L}^* \leq \mathcal{L}_t - \mathcal{L}^* - \gamma_t \left(g_t^{(x)} - g_t^{(y)} \right) + \gamma_t^2 \frac{L\|\boldsymbol{d}^{(t)}\|^2}{2}. \tag{10}$$

Unfortunately, the quantity g_t^{FW} does *not* appear above and we therefore cannot control the oscillation of the sequence (the quantity $g_t^{(x)} - g_t^{(y)}$ can make the sequence increase or decrease). Instead, we must focus on more specific SP optimization settings and introduce other quantities of interest in order to establish convergence.

The asymmetry of the SP. Hammond (1984, p. 165) showed the divergence of the SP-FW algorithm with an extended line-search step-size on some bilinear objectives. She mentioned that the difficulty for SP optimization is contained in this bilinear coupling between $\boldsymbol{x}$ and $\boldsymbol{y}$. More generally, most of the examples of SP functions cited in the introduction can be written in the form:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}) + \boldsymbol{x}^\top M \boldsymbol{y} - g(\boldsymbol{y}), \ f \text{ and } g \text{ convex}. \tag{11}$$

In this setting, the bilinear part M is the only term preventing us to apply theorems on standard FW. Hammond (1984, p. 175) also conjectured that the SP-FW algorithm with $\gamma_t = 1/(t+1)$ performed on an uniformly strongly convex-concave objective function (see (12)) over a polytope should converge. We give a partial answer to this conjecture in the following section.

3 SP-FW for strongly convex functions

Uniform strong convex-concavity. In this section, we will assume that $\mathcal{L}$ is uniformly $(\mu_{\mathcal{X}}, \mu_{\mathcal{Y}})$-strongly convex-concave, which means that the following function is convex-concave:

$$(\boldsymbol{x}, \boldsymbol{y}) \mapsto \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) - \frac{\mu_{\mathcal{X}}}{2}\|\boldsymbol{x}\|^2 + \frac{\mu_{\mathcal{Y}}}{2}\|\boldsymbol{y}\|^2. \tag{12}$$

A new merit function. To prove our theorem, we use a different quantity w_t which is smaller than h_t but still a valid merit function in the case of *strongly convex-concave* SPs (where $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ is thus unique); see (14) below. For $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ a solution of (1), we define the non-negative quantity w_t:

$$w_t := \underbrace{\mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*) - \mathcal{L}^*}_{:= w_t^{(x)}} + \underbrace{\mathcal{L}^* - \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}^{(t)})}_{:= w_t^{(y)}}. \tag{13}$$

Notice that $w_t^{(x)}$ and $w_t^{(y)}$ are non-negative, and that $w_t \leq h_t$ since:

$$\mathcal{L}(\boldsymbol{x}^{(t)}, \widehat{\boldsymbol{y}}^{(t)}) - \mathcal{L}(\widehat{\boldsymbol{x}}^{(t)}, \boldsymbol{y}^{(t)}) \geq \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}^{(t)}).$$

In general, w_t can be zero even if we have not reached a solution. For example, with $\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x} \cdot \boldsymbol{y}$ and $\mathcal{X} = \mathcal{Y} = [-1, 1]$, then $\boldsymbol{x}^* = \boldsymbol{y}^* = \boldsymbol{0}$, implying $w_t = 0$ for any $(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})$. But for a uniformly strongly convex-concave $\mathcal{L}$, this cannot happen and we can prove that w_t has the following nice property (Proposition 15 in Appendix B.6):

$$h_t \leq \sqrt{2} P_{\mathcal{L}} \sqrt{w_t}\,, \tag{14}$$

where

$$P_{\mathcal{L}} \leq \sqrt{2} \sup_{\boldsymbol{z} \in \mathcal{X} \times \mathcal{Y}} \left\{ \frac{\|\nabla_x \mathcal{L}(\boldsymbol{z})\|_{\mathcal{X}^*}}{\sqrt{\mu_{\mathcal{X}}}}, \frac{\|\nabla_y \mathcal{L}(\boldsymbol{z})\|_{\mathcal{Y}^*}}{\sqrt{\mu_{\mathcal{Y}}}} \right\}. \tag{15}$$

Pyramidal width and distance to the border. We now provide a theorem that establishes convergence in two situations: (I) when the SP belongs to the interior of $\mathcal{X} \times \mathcal{Y}$; (P) when the set is a polytope, i.e. when there exist two finite sets such that $\mathcal{X} = \mathrm{conv}(\mathcal{A})$ and $\mathcal{Y} = \mathrm{conv}(\mathcal{B})$). Our convergence result holds when (roughly) the strong convex-concavity of $\mathcal{L}$ is big enough

in comparison to the cross Lipschitz constants L_{XY}, L_{YX} of $\nabla\mathcal{L}$ (defined in (20) below) multiplied by geometric "condition numbers" of each set. The condition number of $\mathcal{X}$ (and similarly for $\mathcal{Y}$) is defined as the ratio of its diameter $D_{\mathcal{X}} := \sup_{\boldsymbol{x},\boldsymbol{x}'\in\mathcal{X}} \|\boldsymbol{x}-\boldsymbol{x}'\|$ over the following appropriate notions of "width":

$$\text{border distance:}\quad \delta_{\mathcal{X}} := \min_{\boldsymbol{s}_x\in\partial\mathcal{X}} \|\boldsymbol{x}^* - \boldsymbol{s}\| \quad \text{for (I)}, \tag{16}$$

$$\text{pyramidal width:}\quad \delta_{\mathcal{A}} := PWidth(\mathcal{A}) \qquad \text{for (P)}. \tag{17}$$

The pyramidal width (17) is formally defined in Eq. 9 of Lacoste-Julien and Jaggi (2015). Given the above constants, we can state below a non-affine invariant version of our convergence theorem (for simplicity). The affine invariant versions of this theorem are given in Thm. 24 and 25 in Appendix D.2 (with proofs).

Theorem 1. *Let $\mathcal{L}$ be a convex-concave function and $\mathcal{X}\times\mathcal{Y}$ a convex and compact set. Assume that the gradient of $\mathcal{L}$ is L-Lipschitz continuous, that $\mathcal{L}$ is $(\mu_{\mathcal{X}},\mu_{\mathcal{Y}})$-strongly convex-concave, and that we are in one of the two following situations:*

(I) *The SP belongs to the interior of $\mathcal{X}\times\mathcal{Y}$. In this case, set $g_t = g_t^{\mathrm{FW}}$ (as in L5 of Alg. 3), $\delta_\mu := \sqrt{\min(\mu_{\mathcal{X}}\delta_{\mathcal{X}}^2, \mu_{\mathcal{Y}}\delta_{\mathcal{Y}}^2)}$ and $a := 1$. "Algorithm" then refers to SP-FW.*

(P) *The sets $\mathcal{X}$ and $\mathcal{Y}$ are polytopes. In this case, set $g_t = g_t^{\mathrm{PFW}}$ (as in L11 of Alg. 3), $\delta_\mu := \sqrt{\min(\mu_{\mathcal{X}}\delta_{\mathcal{A}}^2, \mu_{\mathcal{Y}}\delta_{\mathcal{B}}^2)}$ and $a := \frac{1}{2}$. "Algorithm" then refers to SP-AFW. Here δ_μ needs to use the Euclidean norm for its defining constants.*

In both cases, if $\nu := a - \frac{\sqrt{2}}{\delta_\mu}\max\left\{\frac{D_{\mathcal{X}}L_{XY}}{\sqrt{\mu_{\mathcal{Y}}}}, \frac{D_{\mathcal{Y}}L_{YX}}{\sqrt{\mu_{\mathcal{X}}}}\right\}$ is positive, then the errors h_t (5) of the iterates of the algorithm with step size $\gamma_t = \min\{\gamma_{\max}, \frac{\nu}{2C}g_t\}$ decrease geometrically as

$$h_t = O\left((1-\rho)^{\frac{k(t)}{2}}\right) \quad \text{and} \quad \min_{s\le t} g_s = O\left((1-\rho)^{\frac{k(t)}{2}}\right)$$

where $\rho := \frac{\nu^2\mu}{2C}$, $C := \frac{LD_{\mathcal{X}}^2 + LD_{\mathcal{Y}}^2}{2}$ and $k(t)$ is the number of non-drop step after t steps (see L9 in Alg. 3). In case (I) we have $k(t) = t$ and in case (P) we have $k(t) \ge t/3$. For both algorithms, if $\delta_\mu > 2\max\left\{\frac{D_{\mathcal{X}}L_{XY}}{\mu_{\mathcal{X}}}, \frac{D_{\mathcal{Y}}L_{YX}}{\mu_{\mathcal{Y}}}\right\}$, we also obtain a sublinear rate with the universal choice $\gamma_t = \min\{\gamma_{\max}, \frac{2}{2+k(t)}\}$. This yields the rates:

$$\min_{s\le t} h_s \le \min_{s\le t} g_s^{\mathrm{FW}} = O\left(\frac{1}{t}\right). \tag{18}$$

Clearly, the sublinear rate seems less interesting than the linear one but has the added convenience that the step size can be set without knowledge of various constants that characterize $\mathcal{L}$. Moreover, it provides a partial answer to the conjecture from Hammond (1984).

Proof sketch. Strong convexity is an essential assumption in our proof; it allows us to relate w_t to how close we are to the optimum. Actually, by $\mu_{\mathcal{Y}}$-strong concavity of $\mathcal{L}(\boldsymbol{x}^*,\cdot)$ we have

$$\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*\| \le \sqrt{\frac{2}{\mu_{\mathcal{Y}}}\left(\mathcal{L}^* - \mathcal{L}(\boldsymbol{x}^*,\boldsymbol{y}^{(t)})\right)} = \sqrt{\frac{2}{\mu_{\mathcal{Y}}}w_t^{(y)}}. \tag{19}$$

Now, recall that we assumed that $\nabla\mathcal{L}$ is Lipschitz continuous. In the following, we will call L the *Lipschitz continuity constant* of $\nabla\mathcal{L}$ and L_{XY} and L_{YX} its (cross) *partial Lipschitz constants*. For all $\boldsymbol{x}$, $\boldsymbol{x}' \in \mathcal{X}$, $\boldsymbol{y}$, $\boldsymbol{y}' \in \mathcal{Y}$, these constants satisfy

$$\begin{aligned}\|\nabla_x\mathcal{L}(\boldsymbol{x},\boldsymbol{y}) - \nabla_x\mathcal{L}(\boldsymbol{x},\boldsymbol{y}')\|_{\mathcal{X}^*} &\le L_{XY}\|\boldsymbol{y}-\boldsymbol{y}'\|_{\mathcal{Y}},\\ \|\nabla_y\mathcal{L}(\boldsymbol{x},\boldsymbol{y}) - \nabla_y\mathcal{L}(\boldsymbol{x}',\boldsymbol{y})\|_{\mathcal{Y}^*} &\le L_{YX}\|\boldsymbol{x}-\boldsymbol{x}'\|_{\mathcal{X}}.\end{aligned} \tag{20}$$

Note that $L_{XY}, L_{YX} \le L$ if $\|(\boldsymbol{x},\boldsymbol{y})\| := \|\boldsymbol{x}\|_{\mathcal{X}} + \|\boldsymbol{y}\|_{\mathcal{Y}}$. Then, using Lipschitz continuity of the gradient,

$$\begin{aligned}\mathcal{L}(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^*) \le \mathcal{L}(\boldsymbol{x}^{(t)},\boldsymbol{y}^*) + \gamma\langle \boldsymbol{d}_x^{(t)}, \nabla_x\mathcal{L}(\boldsymbol{x}^{(t)},\boldsymbol{y}^*)\rangle\\ + \gamma^2\frac{L\|\boldsymbol{d}_x^{(t)}\|^2}{2}.\end{aligned} \tag{21}$$

Furthermore, setting $(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x}^{(t)},\boldsymbol{y}^*)$ and $\boldsymbol{y}' = \boldsymbol{y}^{(t)}$ in Equation (20), we have

$$\begin{aligned}w_{t+1}^{(x)} \le w_t^{(x)} - \gamma g_t^{(x)} + \gamma D_{\mathcal{X}}L_{XY}\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*\|\\ + \gamma^2\frac{LD_{\mathcal{X}}^2}{2}.\end{aligned} \tag{22}$$

Finally, combining (22) and (19), we get

$$\begin{aligned}w_{t+1}^{(x)} \le w_t^{(x)} - \gamma g_t^{(x)} + \gamma D_{\mathcal{X}}L_{XY}\sqrt{\frac{2}{\mu_{\mathcal{Y}}}}\sqrt{w_t^{(y)}}\\ + \gamma^2\frac{LD_{\mathcal{X}}^2}{2}.\end{aligned} \tag{23}$$

A similar argument on $-\mathcal{L}(\boldsymbol{x}^*,\boldsymbol{y}^{(t+1)})$ gives a bound on $w_t^{(y)}$ much like (23). Summing both yields:

$$\begin{aligned}w_{t+1} \le w_t - \gamma g_t + 2\gamma\max\left\{\frac{D_{\mathcal{X}}L_{XY}}{\sqrt{\mu_{\mathcal{Y}}}}, \frac{D_{\mathcal{Y}}L_{YX}}{\sqrt{\mu_{\mathcal{X}}}}\right\}\sqrt{w_t}\\ + \gamma^2\frac{LD_{\mathcal{X}}^2 + LD_{\mathcal{Y}}^2}{2}.\end{aligned} \tag{24}$$

We now apply recent developments in the convergence theory of FW methods for strongly convex objectives. Lacoste-Julien and Jaggi (2015) crucially upper bound the square root of the suboptimality error on a convex function with the FW gap if the optimum is in the interior, or with the PFW gap if the set is a polytope (Lemma 18 in Appendix C.2). We continue our proof sketch for case (I) only[1]:

$$\begin{aligned}2\mu_{\mathcal{X}}\delta_{\mathcal{X}}^2\left(\mathcal{L}(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)}) - \mathcal{L}(\boldsymbol{x}^*,\boldsymbol{y}^{(t)})\right) \le \left(g_t^{(x)}\right)^2\\ \text{where}\quad \delta_{\mathcal{X}} := \min_{x\in\partial\mathcal{X}}\|\boldsymbol{x}^* - \boldsymbol{s}\|.\end{aligned} \tag{25}$$

[1] The idea is similar for case (P), but with the additional complication of possible drop steps.

We can also get the respective equation on $\boldsymbol{y}$ with $\delta_{\mathcal{Y}} := \min_{\boldsymbol{y} \in \partial \mathcal{Y}} \|\boldsymbol{y}^* - \boldsymbol{y}\|$ and sum it with the previous one (25) to get:

$$\delta_\mu \sqrt{2 w_t} \leq g_t \text{ where } \delta_\mu := \sqrt{\min(\mu_{\mathcal{X}} \delta_{\mathcal{X}}^2, \mu_{\mathcal{Y}} \delta_{\mathcal{Y}}^2)}. \quad (26)$$

Plugging this last equation into (23) gives us

$$\begin{aligned} w_{t+1} \leq w_t - \nu \gamma g_t + \gamma^2 C \text{ where } C := \frac{L D_{\mathcal{X}}^2 + L D_{\mathcal{Y}}^2}{2} \\ \text{and } \nu := 1 - \frac{\sqrt{2}}{\delta_\mu} \max\left\{ \frac{D_{\mathcal{X}} L_{XY}}{\sqrt{\mu_{\mathcal{Y}}}}, \frac{D_{\mathcal{Y}} L_{YX}}{\sqrt{\mu_{\mathcal{X}}}} \right\}. \end{aligned} \quad (27)$$

The recurrence (27) is typical in the FW literature. We can re-apply standard techniques on the sequence w_t to get a sublinear rate with $\gamma_t = \frac{2}{2+t}$, or a linear rate with $\gamma_t = \min\left\{\gamma_{\max}, \frac{\nu g_t}{2C}\right\}$ (which minimizes the RHS of (27) and actually guarantees that w_t will be *decreasing*). Finally, thanks to strong convexity, a rate on w_t gives us a rate on h_t (by (14)). □

4 SP-FW with strongly convex sets

Strongly convex set. One can (roughly) define strongly convex set as sublevel sets of strongly convex functions (Vial, 1983, Prop. 4.14). In this section, we replace the strong convex-concavity assumption on $\mathcal{L}$ with the assumption that $\mathcal{X}$ and $\mathcal{Y}$ are β-strongly convex *sets*.

Definition 2 (Vial (1983); Polyak (1966))**.** *A convex set $\mathcal{X}$ is said to be β-strongly convex with respect to $\|.\|$ if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ and any $\gamma \in [0,1]$, $B_\beta(\gamma, \boldsymbol{x}, \boldsymbol{y}) \subset \mathcal{X}$ where $B_\beta(\gamma, \boldsymbol{x}, \boldsymbol{y})$ is the $\|.\|$-ball of radius $\gamma(1-\gamma)\frac{\beta}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$ centered at $\gamma \boldsymbol{x} + (1-\gamma)\boldsymbol{y}$.*

Frank-Wolfe for convex optimization over strongly convex sets has been studied by Levitin and Polyak (1966); Demyanov and Rubinov (1970) and Dunn (1979), amongst others. They all obtained a linear rate for the FW algorithm if the norm of the gradient is lower bounded by a constant. More recently, Garber and Hazan (2015) proved a sublinear rate $O(1/t^2)$ by replacing the lower bound on the gradient by a strong convexity assumption on the function. In the VIP setting (3), the linear convergence has been proved if the optimization is done under a strongly convex set but this assumption does *not* extend to $\mathcal{X} \times \mathcal{Y}$ which *cannot* be strongly convex if $\mathcal{X}$ or $\mathcal{Y}$ is not reduced to a single element. In order to prove the convergence, we first prove the Lipschitz continuity of the *FW-corner* function $\boldsymbol{s}(\cdot)$ defined below. A proof of this theorem is given in Appendix E.

Theorem 3. *Let $\mathcal{X}$ and $\mathcal{Y}$ be β-strongly convex sets. If $\min(\|\nabla_x L(\boldsymbol{z})\|_{\mathcal{X}^*}, \|\nabla_y L(\boldsymbol{z})\|_{\mathcal{Y}^*}) \geq \delta > 0$ for all $\boldsymbol{z} \in \mathcal{X} \times \mathcal{Y}$, then the oracle function $\boldsymbol{z} \mapsto \boldsymbol{s}(\boldsymbol{z}) := \arg\min_{\boldsymbol{s} \in \mathcal{X} \times \mathcal{Y}} \langle \boldsymbol{s}, \boldsymbol{r}(\boldsymbol{z}) \rangle$ is well defined and is $\frac{4L}{\delta\beta}$-Lipschitz continuous (using the norm $\|(\boldsymbol{x}, \boldsymbol{y})\|_{\mathcal{X} \times \mathcal{Y}} := \|\boldsymbol{x}\|_{\mathcal{X}} + \|\boldsymbol{y}\|_{\mathcal{Y}}$), where $\boldsymbol{r}(\boldsymbol{z}) := (\nabla_x \mathcal{L}(\boldsymbol{z}), -\nabla_y \mathcal{L}(\boldsymbol{z}))$.*

Convergence rate. When the FW-corner function $\boldsymbol{s}(\cdot)$ is Lipschitz continuous (by Theorem 3), we can actually show that the FW gap is decreasing in the FW direction and get a similar inequality as the standard FW one (8), but, in this case, on the *gaps*: $g_{t+1} \leq g_t(1-\gamma_t) + \gamma_t^2 \|\boldsymbol{s}^{(t)} - \boldsymbol{z}^{(t)}\|^2 C_\delta$. Moreover, one can show that the FW gap on a strongly convex set $\mathcal{X}$ can be lower-bounded by $\|\boldsymbol{s}_x^{(t)} - \boldsymbol{x}^{(t)}\|^2$ (Lemma 27 in Appendix E), by using the fact that $\mathcal{X}$ contains a ball of sufficient radius around the midpoint between $\boldsymbol{s}_x^{(t)}$ and $\boldsymbol{x}^{(t)}$. From these two facts, we can prove the following linear rate of convergence (*not* requiring any *strong* convex-concavity of $\mathcal{L}$).

Theorem 4. *Let $\mathcal{L}$ be a convex-concave function and $\mathcal{X}$ and $\mathcal{Y}$ two compact β-strongly convex sets. Assume that the gradient of $\mathcal{L}$ is L-Lipschitz continuous and that there exists $\delta > 0$ such that $\min(\|\nabla_x L(\boldsymbol{z})\|_*, \|\nabla_y L(\boldsymbol{z})\|_*) \geq \delta \ \forall \boldsymbol{z} \in \mathcal{X} \times \mathcal{Y}$. Set $C_\delta := 2L + \frac{8L^2}{\beta\delta}$. Then the gap g_t^{FW} (7) of the SP-FW algorithm with step size $\gamma_t = \frac{g_t^{\text{FW}}}{\|\boldsymbol{s}^{(t)} - \boldsymbol{z}^{(t)}\|^2 C_\delta}$ converges linearly as $g_t^{\text{FW}} \leq g_0 (1-\rho)^t$, where $\rho := \frac{\beta\delta}{16 C_\delta}$.*

5 SP-FW in the bilinear setting

Fictitious play. In her thesis, Hammond (1984, § 4.3.1) pointed out that for the bilinear setting:

$$\min_{\boldsymbol{x} \in \Delta_p} \max_{\boldsymbol{y} \in \Delta_q} \boldsymbol{x}^\top M \boldsymbol{y} \quad (28)$$

where Δ_p is the probability simplex on p elements, the SP-FW algorithm with step size $\gamma_t = 1/(1+t)$ is equivalent to the fictitious play (FP) algorithm introduced by Brown (1951). The FP algorithm has been widely studied in the game literature. Its convergence has been proved by Robinson (1951), while Shapiro (1958) showed that one can deduce from Robinson's proof a $O(t^{-1/(p+q-2)})$ rate. Around the same time, Karlin (1960) conjectured that the FP algorithm converged at the better rate of $O(t^{-1/2})$, though this conjecture is still open and Shapiro's rate is the only one we are aware of. Interestingly, Daskalakis and Pan (2014) recently showed that Shapiro's rate is also a lower bound if the tie breaking rule gets the worst pick an infinite number of times. Nevertheless, this kind of adversarial tie breaking rule does not seems realistic since this rule is a priori defined by the programmer. In practical cases (by setting a fixed prior order for ties or picking randomly for example), Karlin's Conjecture (Karlin, 1960) is still open. Moreover, we always observed an empirical rate of at least $O(t^{-1/2})$ during our experiments, we thus believe the conjecture to be true for realistic tie breaking rules.

Rate for SP-FW. Via the affine invariance of the FW algorithm and the fact that every polytope with

p vertices is the affine transformation of a probability simplex of dimension p, any rate for the fictitious play algorithm implies a rate for SP-FW.

Corollary 5. *For polytopes $\mathcal{X}$ and $\mathcal{Y}$ with p and q vertices respectively and $\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top M \boldsymbol{y}$, the SP-FW algorithm with step size $\gamma_t = \frac{1}{t+1}$ converges at the rate $h_t = O\left(t^{-\frac{1}{p+q-2}}\right)$.*

This (very slow) convergence rate is mainly of theoretical interest, providing a safety check that the algorithm actually converges. Moreover, if Karlin's strong conjecture is true, we can get a $O(1/\sqrt{t})$ worst case rate which is confirmed by our experiments.

6 Experiments

Toy experiments. First, we test the empirical convergence of our algorithms on a simple saddle point problem over the unit cube in dimension d (whose pyramidal width has the explicit value $1/\sqrt{d}$ by Lemma 4 from Lacoste-Julien and Jaggi (2015)). Thus $\mathcal{X} = \mathcal{Y} := [0,1]^d$ and the linear minimization oracle is simply $\mathrm{LMO}(\cdot) = -0.5 \cdot (\mathrm{sign}(\cdot) - \mathbf{1})$. We consider the following objective function:

$$\frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{x}^*\|_2^2+(\boldsymbol{x}-\boldsymbol{x}^*)^\top M(\boldsymbol{y}-\boldsymbol{y}^*)-\frac{\mu}{2}\|\boldsymbol{y}-\boldsymbol{y}^*\|_2^2 \quad (29)$$

for which we can control the location of the saddle point $(\boldsymbol{x}^*, \boldsymbol{y}^*) \in \mathcal{X} \times \mathcal{Y}$. We generate a matrix M randomly as $M \sim \mathcal{U}([-0.1, 0.1]^{d \times d})$ and keep it fixed for all experiments. For the interior point setup (I), we set $(\boldsymbol{x}^*, \boldsymbol{y}^*) \sim \mathcal{U}([0.25, 0.75]^{2d})$, while we set $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$ to some fixed random vertex of the unit cube for the setup (P). With all these parameters fixed, the constant ν is a function of μ only. We thus vary the strong convexity parameter μ to test various ν's.

We verify the linear convergence expected for the SP-FW algorithm for case (I) in Figure 1a, and for the SP-AFW algorithm for case (P) in Figure 1b. As the adaptive step size (and rate) depends linearly on ν, the linear rate becomes quite slow for small ν. In this regime (in red), the step size $2/(2 + k(t))$ (in orange) can actually perform better, despite its theoretical sublinear rate.

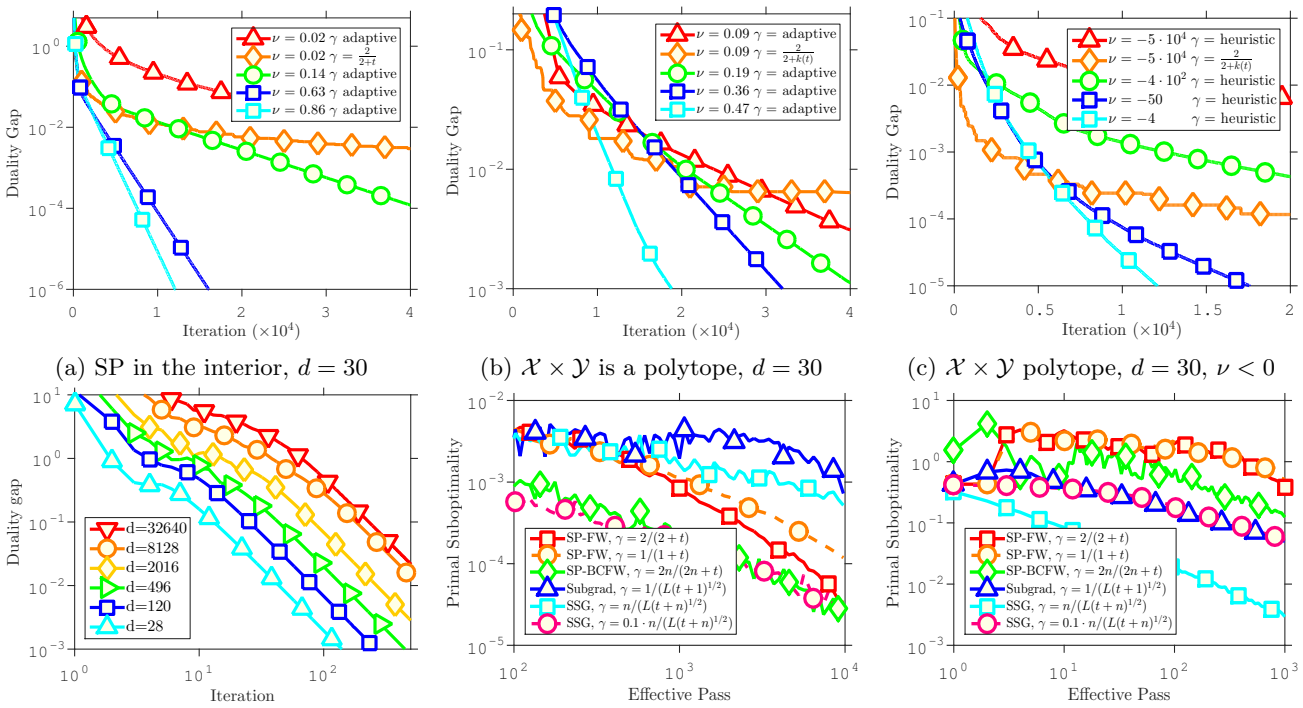
Finally, figure 1c shows that we can observe a linear convergence of SP-AFW even if ν is negative by using a different step size. In this case, we use the heuristic adaptive step size $\gamma_t := g_t/\tilde{C}$ where $\tilde{C} := LD_{\mathcal{X}}^2 + LD_{\mathcal{Y}}^2 + L_{XY}L_{YX}\left(D_{\mathcal{X}}^2/\mu_{\mathcal{X}} + D_{\mathcal{Y}}^2/\mu_{\mathcal{Y}}\right)$. Here $\tilde{C}$ takes into account the coupling between the concave and the convex variable and is motivated from a different proof of convergence that we were not able to complete. The empirical linear convergence in this case is not yet supported by a complete analysis, highlighting the need for more sophisticated arguments.

Graphical games. We now consider a bilinear objective $\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top M \boldsymbol{y}$ where exact projections on the sets is intractable, but we have a tractable LMO. The problem is motivated from the following setup. We consider a game between two universities (A and B) that are admitting s students and have to assign pairs of students into dorms. If students are unhappy with their dorm assignments, they will go to the other university. The game has a payoff matrix M belonging to $\mathbb{R}^{(s(s-1)/2)^2}$ where $M_{ij,kl}$ is the expected tuition that B gets (or A gives up) if A pairs student i with j and B pairs student k with l. Here the actions $\boldsymbol{x}$ and $\boldsymbol{y}$ are both in the marginal polytope of all perfect unipartite matchings. Assume that we are given a graph $G = (V, E)$ with vertices V and edges E. For a subset of nodes $S \subseteq V$, let the induced subgraph $G(S) = (S, E(S))$. Edmonds (1965) showed that any subgraph forming a triangle can contain at most one edge of any perfect matching. This forms an exponential set of linear equalities which define the matching polytope $\mathcal{P}(G) \subset \mathbb{R}^E$ as

$$\{\boldsymbol{x} \,|\, \boldsymbol{x}_e \geq 0, \sum_{e \in E(S)} \boldsymbol{x}_e \leq k, \forall S \subseteq V, |S| = 2k+1, \forall e \in E\}. \quad (30)$$

While this strategy space seems daunting, the LMO can be solved in $\mathcal{O}(s^3)$ time using the blossom algorithm (Edmonds, 1965). We run the SP-FW algorithm with $\gamma_t = 2/(t+2)$ on this problem with $s = 2^j$ students for $j = 3, \ldots, 8$ with results given in Figure 1d ($d = s(s-1)/2$ in the legend represents the dimensionality of the $\boldsymbol{x}$ and $\boldsymbol{y}$ variables). The order of the complexity of the LMO is then $O(d^{3/2})$. In Figure 1d, the observed empirical rate of the SP-FW algorithm (using $\gamma_t = 2/(t+2)$) is $O(1/t^2)$. Empirically, faster rates seem to arise if the solution is at a corner (a pure equilibrium, to be expected for random payoff matrices in light of (Bárány et al., 2007)).

Sparse structured SVM. We finally consider a challenging optimization problem arising from structured prediction. We consider the saddle point formulation (Taskar et al., 2006) for a ℓ_1-regularized structured SVM objective that minimizes the primal cost function $p(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^n \tilde{H}_i(\boldsymbol{w})$, where $\tilde{H}_i(\boldsymbol{w}) = \max_{\boldsymbol{y} \in \mathcal{Y}_i} L_i(\boldsymbol{y}) - \langle \boldsymbol{w}, \boldsymbol{\psi}_i(\boldsymbol{y}) \rangle$ is the structured hinge loss (using the notation from Lacoste-Julien et al. (2013)). We only assume access to the linear oracle computing $\tilde{H}_i(\boldsymbol{w})$. Let M_i have $\left(\boldsymbol{\psi}_i(\boldsymbol{y})\right)_{\boldsymbol{y} \in \mathcal{Y}_i}$ as columns. We can rewrite the minimization problem as a bilinear saddle point problem:



(a) SP in the interior, $d = 30$ (b) $\mathcal{X} \times \mathcal{Y}$ is a polytope, $d = 30$ (c) $\mathcal{X} \times \mathcal{Y}$ polytope, $d = 30$, $\nu < 0$

(d) Graphical games (e) OCR dataset, $R = 0.01$. (f) OCR dataset, $R = 5$.

Figure 1: On Figure 1a, 1b and 1c, we plot on a semilog scale the best gap observed $\min_{i \leq t} g_i$ as a function of t. For experiments 1d, 1e and 1f, the objective function is bilinear and the convergence is sublinear. We give more details about these experiments in Appendix F.

$$\min_{\|\boldsymbol{w}\|_1 \leq R} \frac{1}{n} \sum_i \Big(\max_{\boldsymbol{y}_i \in \mathcal{Y}_i} \boldsymbol{L}_i^\top \boldsymbol{y}_i - \boldsymbol{w}^\top M_i \boldsymbol{y}_i \Big) = \min_{\|\boldsymbol{w}\|_1 \leq R} \frac{1}{n} \sum_i \Big(\max_{\boldsymbol{\alpha}_i \in \Delta(|\mathcal{Y}_i|)} \boldsymbol{L}_i^\top \boldsymbol{\alpha}_i - \boldsymbol{w}^\top M_i \boldsymbol{\alpha}_i \Big). \tag{31}$$

Projecting onto $\Delta(|\mathcal{Y}_i|)$ is normally intractable as the size of $|\mathcal{Y}_i|$ is exponential, but the linear oracle is tractable by assumption. We performed experiments with 100 examples from the OCR dataset ($d_\omega = 4028$) (Taskar et al., 2003). We encoded the structure $\mathcal{Y}_i$ of the i^{th} word with a Markov model: its k^{th} character $\mathcal{Y}_i^{(k)}$ only depends on $\mathcal{Y}_i^{k-1}$ and $\mathcal{Y}_i^{k+1}$. In this case, the oracle function is simply the Viterbi algorithm Viterbi (1967). The average length of a word is approximately 8, hence the dimension of $\mathcal{Y}_i$ is $d_{\mathcal{Y}_i} \approx 26^2 \cdot 8 = 5408$ leading to a large dimension for $\mathcal{Y}$, $d_{\mathcal{Y}} := \sum_{i=1}^n d_{\mathcal{Y}_i} \approx 5 \cdot 10^5$. We run the SP-FW algorithm with step size $\gamma_t = 1/(1+t)$ for which we have a convergence proof (Corollary 5), and with $\gamma_t = 2/(2+t)$, which normally gives better results for FW optimization. We compare with the projected subgradient method (projecting on the ℓ_1-ball is tractable here) with step size $O(1/\sqrt{t})$ (the subgradient of $\tilde{H}_i(\boldsymbol{w})$ is $-\boldsymbol{\psi}_i(\boldsymbol{y}_i^*)$). Following Lacoste-Julien et al. (2013), we also implement a block-coordinate (SP-BCFW) version of SP-FW and compare it with the stochastic projected subgradient method (SSG). As some of the algorithms only work on the primal and to make our result comparable to Lacoste-Julien et al. (2013), we choose to plot the primal suboptimality error $p(\boldsymbol{w}_t) - p^*$ for the different algorithms in Figure 1e and 1f (the $\boldsymbol{\alpha}_t$ iterates for the SP approaches are thus ignored in this error). The performance of SP-BCFW is similar to SSG when we regularize the learning problem heavily (Figure 1e). However, under lower regularization (Figure 1f), SSG (with the correct step size scaling) is faster. This is consistent with the fact that $\boldsymbol{\alpha_t} \neq \boldsymbol{\alpha}^*$ implies larger errors on the primal suboptimality for the SP methods, but we note that an advantage of the SP-FW approach is that the scale of the step size is automatically chosen.

Conclusion. We proposed FW-style algorithms for saddle-point optimization with the same attractive properties as FW, in particular only requiring access to a LMO. We gave the first convergence result for a FW-style algorithm towards a saddle point over polytopes by building on the recent developments on the linear convergence analysis of AFW. However, our experiments let us believe that the condition $\nu > 0$ is not required for the convergence of FW-style algorithms. We thus conjecture that a refined convergence analysis could yield a linear rate for the general uniformly strongly convex-concave functions in both cases (I) and (P), paving the way for further theoretical work.

Acknowledgments

Thanks to N. Ruozzi and A. Benchaouine for helpful discussions. Work supported in part by DARPA N66001-15-2-4026, N66001-15-C-4032 and NSF III-1526914, IIS-1451500, CCF-1302269.

References

A. Ahmadinejad, S. Dehghani, Hajiaghayi, B. Lucier, H. Mahini, and S. Seddighin. From duels to battlefields: Computing equilibria of Blotto and other games. In *AAAI*, 2016.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 2012.

F. Barahona. On the computational complexity of Ising spin glass models. *J. Phys. A: Math. Gen.*, 1982.

I. Bárány, S. Vempala, and A. Vetta. Nash equilibria in random games. *Random Structures & Algorithms*, 31(4):391–405, 2007.

G. Brown. Iterative solution of games by fictitious play. *Activity analysis of production & allocation*, 1951.

B. Cox, A. Juditsky, and A. Nemirovski. Decomposition techniques for bilinear saddle point problems and variational inequalities with affine monotone operators on domains given by linear minimization oracles. *arXiv preprint arXiv:1506.02444*, 2015.

C. Daskalakis and Q. Pan. A counter-example to Karlin's strong conjecture for fictitious play. In *FOCS*, 2014.

V. F. Demyanov and A. M. Rubinov. *Approximate methods in optimization problems*. Elsevier, 1970.

J. C. Dunn. Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals. *SIAM Journal on Control and Optimization*, 1979.

J. Edmonds. Paths, trees and flowers. *Canadian Journal of Mathematics*, 1965.

M. Frank and P. Wolfe. An algorithm for quadratic programming. *Nav. Res. Logist. Q.*, 1956.

D. Garber and E. Hazan. Faster rates for the Frank-Wolfe method over strongly-convex sets. In *ICML*, 2015.

J. H. Hammond. *Solving asymmetric variational inequality problems and systems of equations with generalized nonlinear programming algorithms*. PhD thesis, Massachusetts Institute of Technology, 1984.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms I: Fundamentals*. Springer, 2013.

M. Jaggi. *Sparse convex optimization methods for machine learning*. PhD thesis, ETH Zürich, 2011.

M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *ICML*, 2013.

A. Juditsky and A. Nemirovski. Solving variational inequalities with monotone operators on domains given by linear minimization oracles. *Mathematical Programming*, 2016.

S. Karlin. Mathematical methods and theory in games, programming and economics, 1960.

J. Kelner, Y. Lee, L. Orrechia, and A. Sidford. An almost-linear-time algorithm for approximate max flow in undirected graphs, and its multicommodity generalizations. In *SODA*, 2014.

S.-J. Kim, A. Magnani, and S. Boyd. Robust Fisher discriminant analysis. In *NIPS*, 2005.

D. Koller, N. Megiddo, and B. Von Stengel. Fast algorithms for finding randomized strategies in game trees. In *STOC*, 1994.

G. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 1976.

S. Lacoste-Julien and M. Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. In *NIPS*, 2015.

S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. In *ICML*, 2013.

G. Lan. The complexity of large-scale convex programming under a linear optimization oracle. *arXiv preprint arXiv:1309.5550*, 2013.

T. Larsson and M. Patriksson. A class of gap functions for variational inequalities. *Math. Prog.*, 1994.

E. S. Levitin and B. T. Polyak. Constrained minimization methods. *USSR Computational mathematics and mathematical physics*, 1966.

A. F. Martins, N. A. Smith, P. M. Aguiar, and M. A. Figueiredo. Structured sparsity in structured prediction. In *EMNLP*, 2011.

Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 2007.

M. Patriksson. *Nonlinear Programming and Variational Inequality Problems: A Unified Approach*. Springer, 1999.

B. T. Polyak. Existence theorems and convergence of minimizing sequences in extremum problems with restrictions. *Soviet Math. Dokl*, 1966.

J. Robinson. An iterative method of solving a game. *Annals of mathematics*, 1951.

H. N. Shapiro. Note on a computation method in the theory of games. *Communications on Pure and Applied Mathematics*, 1958.

B. Taskar, C. Guestrin, and D. Koller. Max-margin markov networks. In *NIPS*, 2003.

B. Taskar, S. Lacoste-Julien, and M. I. Jordan. Structured prediction, dual extragradient and Bregman projections. *Journal of Machine Learning Research*, 2006.

J.-P. Vial. Strong and weak convexity of sets and functions. *Mathematics of Operations Research*, 1983.

A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269, 1967.

J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton press, 1944.

N. Xiu and J. Zhang. Some recent advances in projection-type methods for variational inequalities. *Journal of Computational and Applied Mathematics*, 2003.

D. L. Zhu and P. Marcotte. Convergence properties of feasible descent methods for solving variational inequalities in banach spaces. *Computational Optimization and Applications*, 1998.

Appendix

Outline. Appendix A provides more details about the saddle point away-step Frank-Wolfe (SP-AFW) algorithm. Appendix B is about the affine invariant formulation of our algorithms, therein, we introduce some affine invariant constants and prove relevant bounds. Appendix C presents some relationships between the primal suboptimalities and dual gaps useful for the convergence proof. Appendix D gives the affine invariant convergence proofs of SP-FW and SP-AFW in the strongly convex function setting introduced in Section 3. Appendix E gives the proof of linear convergence of SP-FW in the strongly convex set setting as defined in Section 4. Finally, Appendix F provides details on the experiments.

A Saddle point away-step Frank-Wolfe (SP-AFW)

In this section, we describe our algorithms SP-AFW and SP-PFW with a main focus on how the away direction is chosen. We also rigorously define a *drop step* and prove an upper bound on their number. In this section, we will assume that there exist two finites sets $\mathcal{A}$ and $\mathcal{B}$ such that $\mathcal{X} = \operatorname{conv}(\mathcal{A})$ and $\mathcal{Y} = \operatorname{conv}(\mathcal{B})$.

Active sets. Our definition of *active set* is an extension of the one provided in Lacoste-Julien and Jaggi (2015), we follow closely their notation and their results. Assume that we have the current expansion,

$$\boldsymbol{x}^{(t)} = \sum_{\boldsymbol{v}_x \in \mathcal{S}_x^{(t)}} \alpha_{\boldsymbol{v}_x}^{(t)} \boldsymbol{v}_x \quad \text{where} \quad \mathcal{S}_x^{(t)} := \left\{ \boldsymbol{v}_x \in \mathcal{A} \,;\, \alpha_{\boldsymbol{v}_x}^{(t)} > 0 \right\}, \tag{32}$$

and a similar one for $\boldsymbol{y}^{(t)}$. Then, the current iterate has a sparse representation as a convex combination of all possible pairs of atoms belonging to $S_x^{(t)}$ and $S_y^{(t)}$, i.e.

$$\boldsymbol{z}^{(t)} = \sum_{\boldsymbol{v} \in \mathcal{S}^{(t)}} \alpha_{\boldsymbol{v}}^{(t)} \boldsymbol{v} \quad \text{where} \quad \mathcal{S}^{(t)} := \left\{ \boldsymbol{v} \in \mathcal{A} \times \mathcal{B} \,;\, \alpha_{\boldsymbol{v}}^{(t)} := \alpha_{\boldsymbol{v}_x}^{(t)} \alpha_{\boldsymbol{v}_y}^{(t)} > 0 \right\}. \tag{33}$$

The set $\mathcal{S}^{(t)}$ is the current (implicit) *active set*. Note that after t iteration, the current iterate $\boldsymbol{z}^{(t)}$ is t-sparse whereas the size of the active set $\mathcal{S}^{(t)}$ defined in (33) can be t^2. Fortunately, we only need to track at most t corners in $\mathcal{A}$ and t ones in $\mathcal{B}$ to get this bigger active set. We can now define the maximal step size for an away direction.

Maximal step size. For the standard AFW algorithm, Lacoste-Julien and Jaggi (2015) suggest to use the maximum step size $\gamma_{\max} = \alpha_{\boldsymbol{v}^{(t)}} / (1 - \alpha_{\boldsymbol{v}^{(t)}})$ when using the away direction $\boldsymbol{z}^{(t)} - \boldsymbol{v}^{(t)}$, to guarantee that the next iterate stays feasible. Because we have a product structure of two blocks, we actually consider more possible away directions by maintaining a separate convex combination on each block in our Algorithm 3 (SP-AFW) and 4 (SP-PFW). More precisely, suppose that we have $\boldsymbol{x}^{(t)} = \sum_{\boldsymbol{v}_x \in S_x^{(t)}} \alpha_{\boldsymbol{v}_x}^{(t)} \boldsymbol{v}_x$ and $\boldsymbol{y}^{(t)} = \sum_{\boldsymbol{v}_y \in S_y^{(t)}} \alpha_{\boldsymbol{v}_y}^{(t)} \boldsymbol{v}_y$, then the following maximum step size $\gamma_{\max}$ (for AFW) ensures that the iterate $\boldsymbol{z}^{(t+1)}$ stays feasible:

$$\boldsymbol{z}^{(t+1)} := \boldsymbol{z}^{(t)} + \gamma_t \boldsymbol{d}_A^{(t)} \quad \text{with} \quad \gamma_t \in [0, \gamma_{\max}] \quad \text{and} \quad \gamma_{\max} := \min \left\{ \frac{\alpha_{\boldsymbol{v}_x^{(t)}}^{(t)}}{1 - \alpha_{\boldsymbol{v}_x^{(t)}}^{(t)}}, \frac{\alpha_{\boldsymbol{v}_y^{(t)}}^{(t)}}{1 - \alpha_{\boldsymbol{v}_y^{(t)}}^{(t)}} \right\}. \tag{34}$$

A larger γ_t makes one of the coefficients in the convex combination for the iterate negative, thus no more guaranteeing that the iterate stays feasible. A similar argument can be used to derive the maximal step size for the PFW direction in Algorithm 4.

Drop steps. A *drop step* is when $\gamma_t = \gamma_{\max}$ for the away-step update (34) (Lacoste-Julien and Jaggi, 2015). In this case, at least one corner is removed from the active set. We show later in Lemma 23 that we can still guarantee progress for this step, i.e. $w_{t+1} < w_t$, but this progress be arbitrarily small since $\gamma_{\max}$ can be arbitrarily small. Lacoste-Julien and Jaggi (2015) shows that the number of drop steps for AFW is at most half of the number of iterations. Because we are maintaining two independent active sets in our formulation, we can obtain more drop steps, but we can still adapt their argument to obtain that the number of drop steps for SP-AFW is at most two thirds the number of iterations (assuming that the algorithm is initialized with only one atom per active set). In the SP-AFW algorithm, either a FW step is jointly made on both blocks, or an away-step is done on both blocks. Let us call A_t the number of FW steps (which potentially adds an atom in $S_x^{(t)}$ and $S_y^{(t)}$) and $D_t^{(x)}$ (resp $D_t^{(y)}$) the number of steps that removed at least one atom from $S_x^{(t)}$ ($S_y^{(t)}$). Finally, we call D_t the number of *drop steps*, i.e., the number of *away steps* where at least one atom from $S_x^{(t)}$ or $S_y^{(t)}$ have been removed (and thus $\gamma_t = \gamma_{\max}$ for these). Because a step is either a FW step or an away step, we have:

$$A_t + D_t \leq t\,. \tag{35}$$

We also have that $D_t^{(x)} + D_t^{(y)} \geq D_t$ by definition of D_t. Because a FW step adds at most one atom in an active set while a drop step removes one, we have (supposing that $|S_x^{(0)}| = |S_y^{(0)}| = 1$):

$$1 + A_t - D_t^{(x)} \geq |S_x^{(t)}| \quad \text{and} \quad 1 + A_t - D_t^{(y)} \geq |S_y^{(t)}|. \tag{36}$$

Adding these two relations, we get:

$$2 + 2A_t \geq |S_x^{(t)}| + |S_y^{(t)}| + D_t^{(x)} + D_t^{(y)} \geq 2 + D_t\,, \tag{37}$$

using the fact that each active set as at least one element. We thus obtain $D_t \leq 2A_t$. Combining with (35), we get:

$$D_t \leq \frac{2}{3}t\,, \tag{38}$$

as claimed.

B Affine invariant formulation of SP-FW

In this section, we define the affine invariant constants of a convex function f and their extension to a convex-concave function $\mathcal{L}$. These constants are important as the FW-type algorithms are affine invariant if their step size are defined using affine invariant quantities. We can upper bound these constants using the non affine invariant constants defined in the main paper. Hence a convergence rate with affine invariant constants will immediately imply a rate with the constant introduced in the main paper.

B.1 The Lipschitz constants

We define the Lipschitz constant L of the gradient of the function f with respect to the norm $\|\cdot\|$ by using a dual pairing of norms, i.e. L is a constant such that

$$\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}, \qquad \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}')\|_* \leq L\|\boldsymbol{x} - \boldsymbol{x}'\|, \tag{39}$$

where $\|\boldsymbol{y}\|_* := \sup_{\boldsymbol{x}\in\mathbb{R}^d, \|\boldsymbol{x}\|\leq 1} \boldsymbol{y}^T\boldsymbol{x}$ is the dual norm of $\|\cdot\|$. For a convex-concave function, we also consider the partial Lipschitz constants with respect to different blocks as follows.

For more generality, we consider the dual pairing of norms $(\|\cdot\|_{\mathcal{X}}, \|\cdot\|_{\mathcal{X}^*})$ on $\mathcal{X}$, and similarly $(\|\cdot\|_{\mathcal{Y}}, \|\cdot\|_{\mathcal{Y}^*})$ on $\mathcal{Y}$. We also define the norm on the product space $\mathcal{X} \times \mathcal{Y}$ as the ℓ_1-norm on the components: $\|(\boldsymbol{x}, \boldsymbol{y})\|_{\mathcal{X}\times\mathcal{Y}} := \|\boldsymbol{x}\|_{\mathcal{X}} + \|\boldsymbol{y}\|_{\mathcal{Y}}$. We thus have that the dual norm of $\mathcal{X} \times \mathcal{Y}$ is the

ℓ_∞-norm of the dual norms: $\|(\boldsymbol{x}, \boldsymbol{y})\|_{(\mathcal{X}\times\mathcal{Y})^*} = \max(\|\boldsymbol{x}\|_{\mathcal{X}^*}, \|\boldsymbol{y}\|_{\mathcal{Y}^*})$. The *partial* Lipschitz constants L_{XX}, L_{YY}, L_{XY} and L_{YX} of the gradient of the function $\mathcal{L}$ with respect to these norms are the constants such that for all $\boldsymbol{x}$, $\boldsymbol{x}' \in \mathcal{X}$ and $\boldsymbol{y}$, $\boldsymbol{y}' \in \mathcal{Y}$,

$$\begin{aligned} \|\nabla_x \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) - \nabla_x \mathcal{L}(\boldsymbol{x}', \boldsymbol{y})\|_{\mathcal{X}^*} \leq L_{XX}\|\boldsymbol{x} - \boldsymbol{x}'\|_{\mathcal{X}}, \quad \|\nabla_y \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) - \nabla_y \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}')\|_{\mathcal{Y}^*} \leq L_{YY}\,\|\boldsymbol{y} - \boldsymbol{y}'\|_{\mathcal{Y}}, \\ \|\nabla_x \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) - \nabla_x \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}')\|_{\mathcal{X}^*} \leq L_{XY}\|\boldsymbol{y} - \boldsymbol{y}'\|_{\mathcal{Y}}, \quad \|\nabla_y \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) - \nabla_y \mathcal{L}(\boldsymbol{x}', \boldsymbol{y})\|_{\mathcal{Y}^*} \leq L_{YX}\,\|\boldsymbol{x} - \boldsymbol{x}'\|_{\mathcal{X}}. \end{aligned} \tag{40}$$

Note that the cross partial Lipschitz constants L_{XY} and L_{YX} do not necessarily use a dual pairing as $\mathcal{X}$ and $\mathcal{Y}$ could be very different spaces. On the other hand, as the possibilities in (40) are special cases of (39) when considering the ℓ_1-norm of this product domain, one can easily deduce that the partial Lipschitz constants can always be taken to be smaller than the full Lipschitz constant for the gradient of $\mathcal{L}$, i.e., we have that $L \geq \max(L_{XX}, L_{XY}, L_{YX}, L_{YY})$.

B.2 The curvature: an affine invariant measure of smoothness

To prove the convergence of the Frank-Wolfe algorithm, the typical affine invariant analysis proof in the FW literature assumes that the curvature of the objective function is bounded, where the curvature is defined by Jaggi (2013) for example. We give below a slight generalization of this curvature notion in order to handle the convergence analysis of FW with away-steps.[2] It has the same upper bound as the traditional curvature constant (see Proposition 6).

Curvature. [Slight generalization of Jaggi (2013)] Let $f : \mathcal{X} \to \mathbb{R}$ be a convex function, we define the curvature C_f of f as

$$C_f := \sup_{\substack{\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{v} \in \mathcal{X}, \\ \gamma > 0 \text{ s.t.} \\ \boldsymbol{x}_\gamma := \boldsymbol{x} + \gamma \boldsymbol{d} \in \mathcal{X} \\ \text{with } \boldsymbol{d} := \boldsymbol{s} - \boldsymbol{v}}} \frac{2}{\gamma^2}(f(\boldsymbol{x}_\gamma) - f(\boldsymbol{x}) - \gamma \langle \boldsymbol{d}, \nabla f(\boldsymbol{x})\rangle). \tag{41}$$

Note that only the *feasible* step sizes γ are considered in the definition of C_f, i.e., γ such that $\boldsymbol{x}_\gamma \in \mathcal{X}$. If the gradient of the objective function is Lipschitz continuous, the curvature is upper bounded.

Proposition 6 (Simple generalization of Lemma 7 in Jaggi (2013))**.** *Let f be a convex and continuously differentiable function on $\mathcal{X}$ with its gradient ∇f L-Lipschitz continuous w.r.t. some norm $\|.\|$ in dual pairing over the domain $\mathcal{X}$. Then*

$$C_f \leq D_{\mathcal{X}}^2 L\,, \tag{42}$$

where $D_{\mathcal{X}} := \sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}} \|\boldsymbol{x} - \boldsymbol{x}'\|$ is the diameter of $\mathcal{X}$.

Lemma 1.2.3 in Nesterov (2004), Jaggi (2013). Let $\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{v} \in \mathcal{X}$, set $\boldsymbol{d} := \boldsymbol{s} - \boldsymbol{v}$ and $\boldsymbol{x}_\gamma = \boldsymbol{x} + \gamma \boldsymbol{d}$ for some $\gamma > 0$ such that $\boldsymbol{x}_\gamma \in \mathcal{X}$. Then by the fundamental theorem of calculus,

$$f(\boldsymbol{x}_\gamma) = f(\boldsymbol{x}) + \int_0^\gamma \langle \boldsymbol{d}, \nabla f(\boldsymbol{x} + t\boldsymbol{d})\rangle\, dt. \tag{43}$$

[2] The change is to consider the more general directions $\boldsymbol{s} - \boldsymbol{v}$ instead of just $\boldsymbol{s} - \boldsymbol{x}$, and also any feasible positive step size. See also Footnote 8 in Lacoste-Julien and Jaggi (2013) for a related discussion. A different (bigger) constant was required in (Lacoste-Julien and Jaggi, 2015) for the analysis of AFW because they used a line-search.

Hence, we can write

$$\begin{aligned} f(\boldsymbol{x}_\gamma) - f(\boldsymbol{x}) - \gamma \langle \boldsymbol{d}, \nabla f(\boldsymbol{x}) \rangle &= \int_0^\gamma \langle \boldsymbol{d}, \nabla f(\boldsymbol{x} + t\boldsymbol{d}) - \nabla f(\boldsymbol{x}) \rangle \, dt \\ &\leq \|\boldsymbol{d}\| \int_0^\gamma \|\nabla f(\boldsymbol{x} + t\boldsymbol{d}) - \nabla f(\boldsymbol{x})\|_* dt \\ &\leq D_{\mathcal{X}}^2 L \int_0^\gamma t dt \\ &\leq \frac{\gamma^2}{2} D_{\mathcal{X}}^2 L. \end{aligned} \tag{44}$$

Thus for all $\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{v} \in \mathcal{X}$ and $\boldsymbol{x}_\gamma = \boldsymbol{x} + \gamma(\boldsymbol{s} - \boldsymbol{v})$ for $\gamma > 0$ such that $\boldsymbol{x}_\gamma \in \mathcal{X}$, we have

$$\frac{2}{\gamma^2}(f(\boldsymbol{x}_\gamma) - f(\boldsymbol{x}) - \gamma \langle \boldsymbol{s} - \boldsymbol{v}, \nabla f(\boldsymbol{x}) \rangle) \leq L D_{\mathcal{X}}^2. \tag{45}$$

The supremum is then upper bounded by the claimed quantity. □

Osokin et al. (2016, Appendix C.1) illustrate well the importance of the affine invariant curvature constant for Frank-Wolfe algorithms in their paragraph titled "Lipschitz and curvature constants". They provide a concrete example where the wrong choice of norm for a specific domain $\mathcal{X}$ can make the upper bound of Proposition 6 extremely loose, and thus practically useless for an analysis.

We will therefore extend the curvature constant to the convex-concave function $\mathcal{L}$ by simply defining it as the maximum of the curvatures of the functions belonging to the family $(\boldsymbol{x}' \mapsto \mathcal{L}(\boldsymbol{x}', \boldsymbol{y}), \boldsymbol{y}' \mapsto -\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}'))_{\boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y}}$ (see Section B.4). But before that, we review affine invariant analogues of the strong convexity constants that will be useful for the analysis.

B.3 Affine invariant measures of strong convexity

In this section, we review two affine invariant measures of strong convexity that were proposed by Lacoste-Julien and Jaggi (2013)(Lacoste-Julien and Jaggi, 2015) for the affine invariant linear convergence analysis of the standard Frank-Wolfe algorithm (using the "interior strong convexity constant") or the away-step Frank-Wolfe algorithm (using the "geometric strong convexity constant"). We will re-use them for the affine invariant analysis of the convergence of SP-FW or SP-AFW algorithms. In a similar way as the curvature constant C_f includes information about the constraint set $\mathcal{X}$ and the Lipschitz continuity of the gradient of f together, these constants both include the information about the constraint set $\mathcal{X}$ and the strong convexity of a function f together.

Interior strong convexity constant. [based on Lacoste-Julien and Jaggi (2013)] Let $\boldsymbol{x}_c$ be a point in the relative interior of $\mathcal{X}$. The *interior strong convexity constant* for f with respect to the reference point $\boldsymbol{x}_c$ is defined as

$$\mu_f^{\boldsymbol{x}_c} := \inf_{\substack{\boldsymbol{x} \in \mathcal{X} \setminus \{\boldsymbol{x}_c\} \\ \boldsymbol{s} = \bar{\boldsymbol{s}}(\boldsymbol{x}, \boldsymbol{x}_c, \mathcal{X}) \\ \gamma \in (0,1], \\ \boldsymbol{z} = \boldsymbol{x} + \gamma(\boldsymbol{s} - \boldsymbol{x})}} \frac{2}{\gamma^2} \left(f(\boldsymbol{z}) - f(\boldsymbol{x}) - \langle \boldsymbol{z} - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle \right). \tag{46}$$

Here, we follow the notation of Lacoste-Julien and Jaggi (2013) and take the point $\boldsymbol{s}$ to be the point where the ray from $\boldsymbol{x}$ to the reference point $\boldsymbol{x}_c$ pinches the boundary of the set $\mathcal{X}$, i.e. $\bar{\boldsymbol{s}}(\boldsymbol{x}, \boldsymbol{x}_c, \mathcal{X}) := \text{ray}(\boldsymbol{x}, \boldsymbol{x}_c) \cap \partial\mathcal{X}$, where $\partial\mathcal{X}$ is the boundary of the convex set $\mathcal{X}$.

We note that in the original definition (Lacoste-Julien and Jaggi, 2013), $\boldsymbol{x}_c$ was the (unique) optimum point for a strongly convex function f over $\mathcal{X}$. The optimality of $\boldsymbol{x}_c$ is actually not needed in the

definition and so we generalize it here to any point $\boldsymbol{x}_c$ in the relative interior of $\mathcal{X}$, as this will be useful in our convergence proof for SP-FW.

For completeness, we include here the important lower bound from (Lacoste-Julien and Jaggi, 2013) on the interior strong convexity constant in terms of the strong convexity of the function f.

Proposition 7 (Lower bound on $\mu^{\boldsymbol{x}_c}$ from Lacoste-Julien and Jaggi (2013, Lemma 2))**.** *Let f be a convex differentiable function and suppose that f is strongly convex w.r.t. to some arbitrary norm $\|\cdot\|$ over the domain $\mathcal{X}$ with strong-convexity constant $\mu_f > 0$. Furthermore, suppose that the reference point $\boldsymbol{x}_c$ lies in the relative interior of $\mathcal{X}$, i.e., $\delta_c := \min_{\boldsymbol{s}\in\partial\mathcal{X}} \|\boldsymbol{s} - \boldsymbol{x}_c\| > 0$. Then the interior strong convexity constant $\mu_f^{\boldsymbol{x}_c}$ (46) is lower bounded as follows:*

$$\mu_f^{\boldsymbol{x}_c} \geq \mu_f \delta_c^2. \tag{47}$$

Proof. Let $\boldsymbol{x}$ and $\boldsymbol{z}$ be defined as in (46), i.e., $\boldsymbol{z} = \boldsymbol{x} + \gamma(\boldsymbol{s} - \boldsymbol{x})$ for some $\gamma > 0$ and where $\boldsymbol{s}$ intersects the boundary of $\mathcal{X}$ with the ray going from $\boldsymbol{x}$ to $\boldsymbol{x}_c$. By the strong convexity of f, we have

$$f(\boldsymbol{z}) - f(\boldsymbol{x}) - \langle \boldsymbol{z} - \boldsymbol{x}, \nabla f(\boldsymbol{x})\rangle \geq \|\boldsymbol{z} - \boldsymbol{x}\|^2 \frac{\mu_f}{2} = \gamma^2 \|\boldsymbol{s} - \boldsymbol{x}\|^2 \frac{\mu_f}{2}. \tag{48}$$

From the definition of $\boldsymbol{s}$, we have that $\boldsymbol{x}_c$ lies between $\boldsymbol{x}$ and $\boldsymbol{s}$ and thus: $\|\boldsymbol{s} - \boldsymbol{x}\| \geq \|\boldsymbol{s} - \boldsymbol{x}_c\| \geq \delta_c$. Combining with (48), we conclude

$$f(\boldsymbol{z}) - f(\boldsymbol{x}) - \langle \boldsymbol{z} - \boldsymbol{x}, \nabla f(\boldsymbol{x})\rangle \geq \gamma^2 \delta_c^2 \frac{\mu_f}{2}, \tag{49}$$

and therefore

$$\mu_f^{\boldsymbol{x}_c} \geq \delta_c^2 \mu_f. \tag{50}$$

□

We now present the affine invariant constant used in the global linear convergence analysis of Frank-Wolfe variants when the convex set $\mathcal{X}$ is a polytope. The *geometric strong convexity constant* was originally introduced by Lacoste-Julien and Jaggi (2013) and (Lacoste-Julien and Jaggi, 2015). To avoid any ambiguity, we will re-use their definitions verbatim in the rest of this section, starting first with a few geometrical definitions and then presenting the affine invariant constant. In these definitions, they assume that a *finite* set $\mathcal{A}$ of vectors (that they call *atoms*) is given such that $\mathcal{X} = \text{conv}(\mathcal{A})$ (which always exists when $\mathcal{X}$ is a polytope).

Directional Width. [Lacoste-Julien and Jaggi (2015)] The *directional width* of a set $\mathcal{A}$ with respect to a direction $\boldsymbol{r}$ is defined as $dirW(\mathcal{A}, \boldsymbol{r}) := \max_{\boldsymbol{s},\boldsymbol{v}\in\mathcal{A}} \left\langle \frac{\boldsymbol{r}}{\|\boldsymbol{r}\|_2}, \boldsymbol{s} - \boldsymbol{v} \right\rangle$. The *width* of $\mathcal{A}$ is the minimum directional width over all possible directions in its affine hull.

Pyramidal Directional Width. [Lacoste-Julien and Jaggi (2015)] We define the *pyramidal directional width* of a set $\mathcal{A}$ with respect to a direction $\boldsymbol{r}$ and a base point $\boldsymbol{x} \in \mathcal{X}$ to be

$$PdirW(\mathcal{A}, \boldsymbol{r}, \boldsymbol{x}) := \min_{\mathcal{S}\in\mathcal{S}_{\boldsymbol{x}}} dirW(\mathcal{S} \cup \{\boldsymbol{s}(\mathcal{A}, \boldsymbol{r})\},\ \boldsymbol{r}) = \min_{\mathcal{S}\in\mathcal{S}_{\boldsymbol{x}}} \max_{\boldsymbol{s}\in\mathcal{A},\boldsymbol{v}\in\mathcal{S}} \left\langle \frac{\boldsymbol{r}}{\|\boldsymbol{r}\|_2}, \boldsymbol{s} - \boldsymbol{v} \right\rangle, \tag{51}$$

where $\mathcal{S}_{\boldsymbol{x}} := \{\mathcal{S} \,|\, \mathcal{S} \subseteq \mathcal{A}$ such that $\boldsymbol{x}$ is a proper[3] convex combination of all the elements in $\mathcal{S}\}$, and $\boldsymbol{s}(\mathcal{A}, \boldsymbol{r}) := \arg\max_{\boldsymbol{v}\in\mathcal{A}} \langle \boldsymbol{r}, \boldsymbol{v}\rangle$ is the FW atom used as a summit, when using the convention in this section that $\boldsymbol{r} := -\nabla f(\boldsymbol{x})$.

[3] By *proper* convex combination, we mean that all coefficients are non-zero in the convex combination.

Pyramidal Width. [Lacoste-Julien and Jaggi (2015)] To define the pyramidal width of a set, we take the minimum over the cone of possible *feasible* directions $\boldsymbol{r}$ (in order to avoid the problem of zero width).
A direction $\boldsymbol{r}$ is *feasible* for $\mathcal{A}$ from $\boldsymbol{x}$ if it points inwards $\operatorname{conv}(\mathcal{A})$, (i.e. $\boldsymbol{r} \in \operatorname{cone}(\mathcal{A} - \boldsymbol{x})$).
We define the *pyramidal width* of a set $\mathcal{A}$ to be the smallest pyramidal width of all its faces, i.e.

$$PWidth(\mathcal{A}) := \min_{\substack{\mathcal{K} \in \text{faces}(\operatorname{conv}(\mathcal{A})) \\ \boldsymbol{x} \in \mathcal{K} \\ \boldsymbol{r} \in \operatorname{cone}(\mathcal{K} - \boldsymbol{x}) \setminus \{\mathbf{0}\}}} PdirW(\mathcal{K} \cap \mathcal{A}, \boldsymbol{r}, \boldsymbol{x}). \tag{52}$$

Geometric strong convexity constant. [Lacoste-Julien and Jaggi (2015)] The *geometric strong convexity constant* of f (over the set of atoms $\mathcal{A}$ which is left implicit) is:

$$\mu_f^{\mathrm{A}} := \inf_{\boldsymbol{x} \in \mathcal{X}} \inf_{\substack{\boldsymbol{x}^* \in \mathcal{X} \\ s.t \ \langle \nabla f(\boldsymbol{x}), \boldsymbol{x}^* - \boldsymbol{x} \rangle < 0}} \frac{2}{\gamma^{\mathrm{A}}(\boldsymbol{x}, \boldsymbol{x}^*)^2} (f(\boldsymbol{x}^*) - f(\boldsymbol{x}) - \langle \boldsymbol{x}^* - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle) \tag{53}$$

where $\gamma^{\mathrm{A}}(\boldsymbol{x}, \boldsymbol{x}^*) := \frac{\langle -\nabla f(\boldsymbol{x}), \boldsymbol{x}^* - \boldsymbol{x} \rangle}{\langle -\nabla f(\boldsymbol{x}), \boldsymbol{s}_f(\boldsymbol{x}) - \boldsymbol{v}_f(\boldsymbol{x}) \rangle}$ and $\mathcal{X} = \operatorname{conv}(\mathcal{A})$. The quantity $\boldsymbol{s}_f(\boldsymbol{x})$ represents the FW corner picked when running the FW algorithm on f when at $\boldsymbol{x}$; while $\boldsymbol{v}_f(\boldsymbol{x})$ represents the worst-case possible away atom that AFW could pick (and this is where the dependence on $\mathcal{A}$ appears). We now define these quantities more precisely. Recall that the set of possible active sets is $\mathcal{S}_{\boldsymbol{x}} := \{\mathcal{S} \,|\, \mathcal{S} \subseteq \mathcal{A}$ such that $\boldsymbol{x}$ is a proper convex combination of all the elements in $\mathcal{S}\}$. For a given set $\mathcal{S}$, we write $\boldsymbol{v}_{\mathcal{S}}(\boldsymbol{x}) := \arg\max_{\boldsymbol{v} \in \mathcal{S}} \langle \nabla f(\boldsymbol{x}), \boldsymbol{v} \rangle$ for the away atom in the algorithm supposing that the current set of active atoms is $\mathcal{S}$. Finally, we define $\boldsymbol{v}_f(\boldsymbol{x}) := \underset{\{\boldsymbol{v} = \boldsymbol{v}_{\mathcal{S}}(\boldsymbol{x}) \,|\, \mathcal{S} \in \mathcal{S}_{\boldsymbol{x}}\}}{\arg\min} \langle \nabla f(\boldsymbol{x}), \boldsymbol{v} \rangle$ to be the worst-case away atom (that is, the atom which would yield the smallest away descent). An important property coming from this definition that we will use later is that for $\boldsymbol{s}^{(t)}$ and $\boldsymbol{v}^{(t)}$ being possible FW and away atoms (respectively) appearing during the AFW algorithm (consider Algorithm 3 ran only on $\mathcal{X}$), then we have:

$$g_t^{\mathrm{PFW}} := \left\langle \boldsymbol{s}^{(t)} - \boldsymbol{v}^{(t)}, -\nabla f(\boldsymbol{x}^{(t)}) \right\rangle \geq \left\langle \boldsymbol{s}_f(\boldsymbol{x}^{(t)}) - \boldsymbol{v}_f(\boldsymbol{x}^{(t)}), -\nabla f(\boldsymbol{x}^{(t)}) \right\rangle. \tag{54}$$

The following important theorem from (Lacoste-Julien and Jaggi, 2015) lower bounds the geometric strong convexity constant of f in terms of both the strong convexity constant of f, as well as the pyramidal width of $\mathcal{X} = \operatorname{conv}(\mathcal{A})$ defined as $PWidth(\mathcal{A})$ (52).

Proposition 8 (Lower bound for μ_f^{A} from Lacoste-Julien and Jaggi (2015, Theorem 6))**.** *Let f be a convex differentiable function and suppose that f is μ-*strongly convex *w.r.t. to the Euclidean norm $\|\cdot\|_2$ over the domain $\mathcal{X} = \operatorname{conv}(\mathcal{A})$ with strong-convexity constant $\mu \geq 0$. Then*

$$\mu_f^{\mathrm{A}} \geq \mu \cdot (PWidth(\mathcal{A}))^2. \tag{55}$$

The pyramidal width (52) is a geometric quantity with a somewhat intricate definition. Its value is still unknown for many sets (though always strictly positive for finite sets), but Lacoste-Julien and Jaggi (2015, Lemma 4) give its value for the unit cube in $\mathbb{R}^d$ as $1/\sqrt{d}$.

B.4 Curvature and interior strong convexity constant for a convex-concave function

In this subsection, we propose simple convex-concave extensions of the definitions of the affine invariant constants defined introduced in the two previous sections.

To define the convex-concave curvature, we introduce the sets $\mathcal{F}$ and $\mathcal{G}$ of the marginal convex functions.

$$\mathcal{F} := \{\boldsymbol{x}' \mapsto \mathcal{L}(\boldsymbol{x}', \boldsymbol{y})\}_{\boldsymbol{y} \in \mathcal{Y}} \quad \text{and} \quad \mathcal{G} := \{\boldsymbol{y}' \mapsto -\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}')\}_{\boldsymbol{x} \in \mathcal{X}}. \tag{56}$$

Let $\mathcal{L} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ a convex-concave function, we define the curvature pair $(C_{\mathcal{L}_x}, C_{\mathcal{L}_y})$ of $\mathcal{L}$ as

$$(C_{\mathcal{L}_x}, C_{\mathcal{L}_y}) := \left(\sup_{f \in \mathcal{F}} C_f, \sup_{g \in \mathcal{G}} C_g \right). \tag{57}$$

and the curvature of $\mathcal{L}$ as

$$C_{\mathcal{L}} := \frac{C_{\mathcal{L}_x} + C_{\mathcal{L}_y}}{2}. \tag{58}$$

An upper bound on this quantity follows directly from the upper bound on the convex case (Lemma 7 of Jaggi (2013), repeated in our Proposition 6) :

Proposition 9. *Let $\mathcal{L} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a differentiable convex-concave function. If $\mathcal{X}$ and $\mathcal{Y}$ are compact and $\nabla \mathcal{L}$ is Lipschitz continuous, then the curvature of $\mathcal{L}$ is bounded by $\frac{1}{2}(L_{XX} D_{\mathcal{X}}^2 + L_{YY} D_{\mathcal{Y}}^2)$, where L_{XX} (resp L_{YY}) is the largest Lipschitz constant respect to $\boldsymbol{x}$ ($\boldsymbol{y}$) of $\boldsymbol{x} \mapsto \nabla_x \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$ ($\boldsymbol{y} \mapsto \nabla_y \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$).*

Proof. Let f in $\mathcal{F}$,

$$C_f \leq Lip(\nabla f) D_{\mathcal{X}}^2 \leq L_{XX} D_{\mathcal{X}}^2. \tag{59}$$

Similarly, let g in $\mathcal{G}$,

$$C_g \leq Lip(\nabla g) D_{\mathcal{Y}}^2 \leq L_{YY} D_{\mathcal{Y}}^2. \tag{60}$$

Consequently,

$$C_{\mathcal{L}} = \frac{1}{2}(\sup_{f \in \mathcal{F}} C_f + \sup_{g \in \mathcal{G}} C_g) \leq \frac{1}{2}(L_{XX} D_{\mathcal{X}}^2 + L_{YY} D_{\mathcal{Y}}^2). \tag{61}$$

Where $D_{\mathcal{X}}$ and $D_{\mathcal{Y}}$ are the respective diameter of $\mathcal{X}$ and $\mathcal{Y}$. □

Note that L_{XX} and L_{YY} are upper bounded by the global Lipschitz constant of $\nabla \mathcal{L}$. Similarly, we define various notions of strong convex-concavity in the following.

Uniform strong convex-concavity constant. The uniform strong convex-concavity constants is defined as

$$(\mu_{\mathcal{X}}, \mu_{\mathcal{Y}}) := \left(\inf_{f \in \mathcal{G}} \mu_f, \inf_{g \in \mathcal{G}} \mu_g \right) \tag{62}$$

where μ_f is the strong convexity constant of f and μ_g the strong convexity of g.

Under some assumptions this quantity is positive.

Proposition 10. *If the second derivative of $\mathcal{L}$ is continuous, $\mathcal{X}$ and $\mathcal{Y}$ are compact and if for all $f \in \mathcal{F} \cup \mathcal{G}$, $\mu_f > 0$, then $\mu_{\mathcal{X}}$ and $\mu_{\mathcal{Y}}$ are positive.*

Proof. Let us introduce $H_x(\boldsymbol{x}, \boldsymbol{y}) := \nabla_x^2 \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$ the Hessian of the function $\boldsymbol{x} \mapsto \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$. We want to show that the smallest eigenvalue is uniformly bounded on $\mathcal{X} \times \mathcal{Y}$. We know that the smallest eigenvalue lower bounds $\mu_{\mathcal{X}}$,

$$\mu_{\mathcal{X}} \geq \inf_{\substack{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y} \\ \|\boldsymbol{u}\|_2 = 1}} \langle \boldsymbol{u}, H_x(\boldsymbol{x}, \boldsymbol{y}) \cdot \boldsymbol{u} \rangle . \tag{63}$$

But $H_x(\cdot)$ is continuous (because $\nabla_x^2 \mathcal{L}(\cdot)$ is continuous by assumption) and then the function $(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{y}) \mapsto \langle \boldsymbol{u}, H_x(\boldsymbol{x}, \boldsymbol{y}) \cdot \boldsymbol{u} \rangle$ is continuous. Hence since $\mathcal{X} \times \mathcal{Y}$ and the unit ball are compact, the infimum is a minimum which can't be 0 by assumption. Hence $\mu_{\mathcal{X}}$ is positive. Doing the same thing with the smallest eigenvalue of $-\nabla_y^2 \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$, we get that $\mu_{\mathcal{Y}} > 0$. □

A common family of saddle point objectives is of the form $f(x) + x^T My - g(y)$. In this case, we get simply that $(\mu_{\mathcal{X}}, \mu_{\mathcal{Y}}) = (\mu_f, \mu_g)$. An equivalent definition for the uniform strong convex-concavity constant is: $\mathcal{L}$ is $(\mu_{\mathcal{X}}, \mu_{\mathcal{Y}})$-uniform strongly convex-concave function if

$$(\boldsymbol{x}, \boldsymbol{y}) \mapsto \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) - \frac{\mu_{\mathcal{X}}}{2}\|\boldsymbol{x}\|^2 + \frac{\mu_{\mathcal{Y}}}{2}\|\boldsymbol{y}\|^2 \tag{64}$$

is convex-concave.

The following proposition relates the distance between the saddle point and the values of the function. It is a direct consequence from the uniform strong convex-concavity definition (62).

Proposition 11. *Let $\mathcal{L}$ be a uniformly strongly convex-concave function and $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ the saddle point of $\mathcal{L}$. Then we have for all $\boldsymbol{x}$ in $\mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}$,*

$$\sqrt{\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^*) - \mathcal{L}^*} \geq \|\boldsymbol{x}^* - \boldsymbol{x}\|\sqrt{\frac{\mu_{\mathcal{X}}}{2}} \quad \text{and} \quad \sqrt{\mathcal{L}^* - \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y})} \geq \|\boldsymbol{y}^* - \boldsymbol{y}\|\sqrt{\frac{\mu_{\mathcal{Y}}}{2}}. \tag{65}$$

Proof. The saddle point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ is the optimal point of the two strongly convex functions $\boldsymbol{x} \mapsto \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^*)$ and the function $\boldsymbol{y} \mapsto -\mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y})$, so we can use the property of strong convexity on each function and the fact that $\mu_{\mathcal{X}}$ lower bounds the strong convexity constant of $\mathcal{L}(\cdot, \boldsymbol{y}^*)$ (and similarly for $\mu_{\mathcal{Y}}$ with $-\mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y})$) as per the definition (62), to get the required conclusion. □

Now we will introduce the uniform strong convex-concavity constants relatively to our saddle point.

Interior strong convex-concavity. The SP-FW interior strong convex-concavity constants (with respect to the reference point $(\boldsymbol{x}_c, \boldsymbol{y}_c)$) are defined as:

$$\left(\mu_{\mathcal{L}}^{\boldsymbol{x}_c}, \mu_{\mathcal{L}}^{\boldsymbol{y}_c}\right) := \left(\inf_{f \in \mathcal{F}} \mu_f^{\boldsymbol{x}_c}, \inf_{g \in \mathcal{G}} \mu_g^{\boldsymbol{y}_c}\right) \tag{66}$$

where $\mu_f^{\boldsymbol{x}_c}$ is the interior strong convexity constant of f w.r.t to the point $\boldsymbol{x}_c$ and $\mu_g^{\boldsymbol{y}_c}$ is the interior strong convexity constant w.r.t to the point $\boldsymbol{y}_c$. The sets $\mathcal{F}$ and $\mathcal{G}$ are defined in (56). We also define the smallest quantity of both (with the reference point $(\boldsymbol{x}_c, \boldsymbol{y}_c)$ implicit):

$$\mu_{\mathcal{L}}^{\text{int}} = \min\{\mu_{\mathcal{L}}^{\boldsymbol{x}_c}, \mu_{\mathcal{L}}^{\boldsymbol{y}_c}\}. \tag{67}$$

We can lower bound this constant by a quantity depending on the uniform strong convexity constant and the distance of the saddle point to the boundary. The propositions on the strong convex-concavity directly follow from the previous definitions and the analogous proposition on the convex case (Proposition 7)

Proposition 12. *Let $\mathcal{L}$ be a convex-concave function. If the reference point $(\boldsymbol{x}_c, \boldsymbol{y}_c)$ belongs to the relative interior of $\mathcal{X} \times \mathcal{Y}$ and if the function $\mathcal{L}$ is strongly convex-concave with a strong convex-concavity constant $\mu > 0$, then $\mu_{\mathcal{L}}^{int}$ is lower bounded away from zero. More precisely, define $\delta_x := \min_{\boldsymbol{s}_x \in \partial\mathcal{X}} \|\boldsymbol{s}_x - \boldsymbol{x}_c\| > 0$ and $\delta_y := \min_{\boldsymbol{s}_y \in \partial\mathcal{Y}} \|\boldsymbol{s}_y - \boldsymbol{y}_c\|$. Then we have,*

$$\mu_{\mathcal{L}}^{\boldsymbol{x}_c} \geq \mu_{\mathcal{X}} \delta_x^2 \qquad \text{and} \qquad \mu_{\mathcal{L}}^{\boldsymbol{y}_c} \geq \mu_{\mathcal{Y}} \delta_y^2. \tag{68}$$

Proof. Using the Proposition 7 we have,

$$\mu_f^{\boldsymbol{x}_c} \geq \mu_f \cdot \delta_x^2 \geq \mu_{\mathcal{X}} \cdot \delta_x^2, \tag{69}$$

and,

$$\mu_g^{\boldsymbol{y}_c} \geq \mu_g \cdot \delta_y^2 \geq \mu_{\mathcal{Y}} \cdot \delta_y^2. \tag{70}$$

□

When the saddle point is not in the interior of the domain, we define next a constant that takes in consideration the geometry of the sets. If the sets are polytopes, then this constant is positive.

Geometric strong convex-concavity. The SP-FW *geometric* strong convex-concavity constants are defined analogously as the interior strong convex-concavity constants,

$$\left(\mu^{\mathrm{A}}_{\mathcal{L}_x}, \mu^{\mathrm{A}}_{\mathcal{L}_y}\right) := \left(\min_{f\in\mathcal{F}} \mu^{\mathrm{A}}_f, \min_{g\in\mathcal{G}} \mu^{\mathrm{A}}_g\right) ; \quad \mu^{\mathrm{A}}_{\mathcal{L}} := \min\left(\mu^{\mathrm{A}}_{\mathcal{L}_x}, \mu^{\mathrm{A}}_{\mathcal{L}_y}\right), \tag{71}$$

where μ^{A}_f is the geometric strong convexity constant of $f \in \mathcal{F}$ (over $\mathcal{A}$) as defined in (53) (and similarly μ^{A}_g is the geometric strong convexity constant of $g \in \mathcal{G}$ over $\mathcal{B}$).

It is straightforward to notice that the lower bound on the geometric strong convexity constant (Proposition 8) can be extended to the geometric strong convex-concavity constants (where $\mu_{\mathcal{X}}$ and $\mu_{\mathcal{Y}}$ are now assumed to be defined with respect to the Euclidean norm):

$$\mu^{\mathrm{A}}_{\mathcal{L}_x} \geq \mu_{\mathcal{X}}\, PWidth(\mathcal{A})^2 \quad \text{and} \quad \mu^{\mathrm{A}}_{\mathcal{L}_y} \geq \mu_{\mathcal{Y}}\, PWidth(\mathcal{B})^2. \tag{72}$$

B.5 The bilinearity coefficient

In our proof, we need to relate the gradient at the point $(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})$ with the one at the point $(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*)$. We can use the Lipschitz continuity of the gradient for this. We define below affine invariant quantities that can upper bound this difference.

Bilinearity coefficients. let $\mathcal{L}$ be a strongly convex-concave function, and let $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ be its unique saddle point. We define the bilinearity coefficients (M_{XY}, M_{YX}) as,

$$M_{XY} = \sup_{\substack{\boldsymbol{y}\in\mathcal{Y} \\ \boldsymbol{x},\boldsymbol{s},\boldsymbol{v}\in\mathcal{X} \\ \boldsymbol{d}=\boldsymbol{s}-\boldsymbol{v}}} \left\langle \boldsymbol{d}, \frac{\nabla_x \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^*) - \nabla_x \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})}{\sqrt{\mathcal{L}^* - \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y})}} \right\rangle \tag{73}$$

and,

$$M_{YX} := \sup_{\substack{\boldsymbol{x}\in\mathcal{X} \\ \boldsymbol{y},\boldsymbol{s},\boldsymbol{v}\in\mathcal{Y} \\ \boldsymbol{d}=\boldsymbol{s}-\boldsymbol{v}}} \left\langle \boldsymbol{d}, \frac{\nabla_y \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) - \nabla_y \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y})}{\sqrt{\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^*) - \mathcal{L}^*}} \right\rangle. \tag{74}$$

We also define the global bilinearity coefficient as

$$M_{\mathcal{L}} := \max\{M_{XY}, M_{YX}\}. \tag{75}$$

We can upper bound these affine invariant constants with the Lipschitz constant of the gradient, the uniform strong convex-concavity constants and the diameters of the sets.

Proposition 13. *If $\mathcal{X}$ and $\mathcal{Y}$ are compact, $\nabla\mathcal{L}$ is Lipschitz continuous and $\mathcal{L}$ is uniformly strongly convex-concave with constants $(\mu_{\mathcal{X}}, \mu_{\mathcal{Y}})$, then*

$$M_{XY} \leq \sqrt{\frac{2}{\mu_{\mathcal{Y}}}} L_{XY} \cdot D_{\mathcal{X}} \quad \textit{and} \quad M_{YX} \leq \sqrt{\frac{2}{\mu_{\mathcal{X}}}} L_{YX} \cdot D_{\mathcal{Y}} \tag{76}$$

where L_{XY} and L_{YX} are the partial *Lipschitz constants defined in Equation (40). The quantity $D_{\mathcal{X}}$ is the diameter of the compact set $\mathcal{X}$ and $D_{\mathcal{Y}}$ is the diameter of $\mathcal{Y}$.*

Proof.

$$
\begin{aligned}
M_{XY} = & \sup_{\substack{\boldsymbol{y} \in \mathcal{Y} \\ \boldsymbol{x}, \boldsymbol{s}, \boldsymbol{v} \in \mathcal{X} \\ \boldsymbol{d} = \boldsymbol{s} - \boldsymbol{v}}} \left\langle \boldsymbol{d}, \frac{\nabla_x \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^*) - \nabla_x \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})}{\sqrt{\mathcal{L}^* - \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y})}} \right\rangle \\
\leq & \sup_{\substack{\boldsymbol{y} \in \mathcal{Y} \\ \boldsymbol{x}, \boldsymbol{s}, \boldsymbol{v} \in \mathcal{X} \\ \boldsymbol{d} = \boldsymbol{s} - \boldsymbol{v}}} \frac{\|\boldsymbol{d}\|_{\mathcal{X}} \|\nabla_x \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^*) - \nabla_x \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})\|_{\mathcal{X}^*}}{\sqrt{\mathcal{L}^* - \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y})}} \\
\leq & \sup_{\substack{\boldsymbol{y} \in \mathcal{Y} \\ \boldsymbol{s}, \boldsymbol{v} \in \mathcal{X} \\ \boldsymbol{d} = \boldsymbol{s} - \boldsymbol{v}}} \frac{\|\boldsymbol{d}\|_{\mathcal{X}} L_{XY} \|\boldsymbol{y}^* - \boldsymbol{y}\|_{\mathcal{Y}}}{\sqrt{\mathcal{L}^* - \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y})}} \\
\leq & \sup_{\boldsymbol{y} \in \mathcal{Y}} D_{\mathcal{X}} L_{XY} \frac{\|\boldsymbol{y}^* - \boldsymbol{y}\|_{\mathcal{Y}}}{\sqrt{\mathcal{L}^* - \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y})}}.
\end{aligned}
$$

Then using the relation between $\|\boldsymbol{y}^* - \boldsymbol{y}\|_{\mathcal{Y}}$ and $\sqrt{\mathcal{L}^* - \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y})}$ due to strong convexity (Proposition 11)

$$
M_{XY} \leq \sqrt{\frac{2}{\mu_{\mathcal{Y}}}} L_{XY} \cdot D_{\mathcal{X}}. \tag{77}
$$

We use a similar argument for M_{YX} which allows us to conclude. □

B.6 Relation between the primal suboptimalities

In this section, we are going to show that if the objective function $\mathcal{L}$ is uniformly strongly convex-concave, then we have a relation between h_t and w_t. First let us introduce affine invariant constants to relate these quantities (in the context of a given saddle point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$):

$$
P_{\mathcal{X}} := \sup_{\boldsymbol{x} \in \mathcal{X}} \frac{\langle \nabla_x \mathcal{L}(\boldsymbol{x}, \hat{\boldsymbol{y}}(\boldsymbol{x})), \boldsymbol{x} - \boldsymbol{x}^* \rangle}{\sqrt{\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}^*)}} \quad \text{and} \quad P_{\mathcal{Y}} := \sup_{\boldsymbol{y} \in \mathcal{Y}} \frac{\langle \nabla_y \mathcal{L}(\hat{\boldsymbol{x}}(\boldsymbol{y}), \boldsymbol{y}), \boldsymbol{y} - \boldsymbol{y}^* \rangle}{\sqrt{\mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y})}}, \tag{78}
$$

where $\hat{\boldsymbol{y}}(\boldsymbol{x}) := \arg\max_{\boldsymbol{y} \in \mathcal{Y}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$ and $\hat{\boldsymbol{x}}(\boldsymbol{y}) := \arg\min_{\boldsymbol{x} \in \mathcal{X}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$. We also define:

$$
P_{\mathcal{L}} := \max\{P_{\mathcal{X}}, P_{\mathcal{Y}}\}. \tag{79}
$$

These constants can be upper bounded by easily computable constants.

Proposition 14. *For any $(\mu_{\mathcal{X}}, \mu_{\mathcal{Y}})$-uniformly convex-concave function $\mathcal{L}$,*

$$
P_{\mathcal{X}} \leq \sqrt{\frac{2}{\mu_{\mathcal{X}}}} \sup_{\boldsymbol{z} \in \mathcal{X} \times \mathcal{Y}} \|\nabla_x \mathcal{L}(\boldsymbol{z})\|_{\mathcal{X}^*} \quad \textit{and} \quad P_{\mathcal{Y}} \leq \sqrt{\frac{2}{\mu_{\mathcal{Y}}}} \sup_{\boldsymbol{z} \in \mathcal{X} \times \mathcal{Y}} \|\nabla_y \mathcal{L}(\boldsymbol{z})\|_{\mathcal{Y}^*}. \tag{80}
$$

Proof. Let us start from the definition of $P_{\mathcal{X}}$, let $\boldsymbol{x} \in \mathcal{X}$,

$$
\begin{aligned}
\frac{\langle \nabla_x \mathcal{L}(\boldsymbol{x}, \hat{\boldsymbol{y}}(\boldsymbol{x})), \boldsymbol{x} - \boldsymbol{x}^* \rangle}{\sqrt{\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}^*)}} &\leq \frac{\|\boldsymbol{x} - \boldsymbol{x}^*\|_{\mathcal{X}} \cdot \sup\left(\|\nabla_x \mathcal{L}(\boldsymbol{z})\|_{\mathcal{X}^*}\right)}{\sqrt{\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}^*)}} \\
&\leq \sqrt{\frac{2}{\mu_{\mathcal{X}}}} \sup_{\boldsymbol{z} \in \mathcal{X} \times \mathcal{Y}} \|\nabla_x \mathcal{L}(\boldsymbol{z})\|_{\mathcal{X}^*} \qquad \text{(by strong convexity.)}
\end{aligned}
$$

The same way we can get

$$
P_{\mathcal{Y}} \leq \sqrt{\frac{2}{\mu_{\mathcal{Y}}}} \sup_{\boldsymbol{z} \in \mathcal{X} \times \mathcal{Y}} \|\nabla_y \mathcal{L}(\boldsymbol{z})\|_{\mathcal{Y}^*}. \tag{81}
$$

It concludes our proof. □

One way to compute an upper bound on the supremum of the gradient is to use any reference point $\bar{\boldsymbol{z}}$ of the set:

$$\forall \bar{z} \in \mathcal{X} \times \mathcal{Y}, \quad \sup_{\boldsymbol{z} \in \mathcal{X} \times \mathcal{Y}} \|\nabla_x \mathcal{L}(\boldsymbol{z})\|_{\mathcal{X}^*} \leq \nabla_x \mathcal{L}(\bar{\boldsymbol{z}}) + L_{XX} D_{\mathcal{X}} + L_{XY} D_{\mathcal{Y}}. \tag{82}$$

We recall that L_{XX} is the largest (with respect to $\boldsymbol{y}$) Lipschitz constant of $\boldsymbol{x} \mapsto \nabla_x \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$. Note that L_{XX} is upper bounded by the global Lipschitz constant of $\nabla \mathcal{L}$. We can compute an upper bound on the supremum of the norm of $\nabla_y \mathcal{L}$ the same way.

With these above defined affine invariant constants, we can finally relate the two primal suboptimalities as $h_t \leq \mathcal{O}(\sqrt{w_t})$.

Proposition 15. *For a $(\mu_{\mathcal{X}}, \mu_{\mathcal{Y}})$-uniformly strongly convex-concave function $\mathcal{L}$,*

$$h_t \leq P_{\mathcal{L}} \sqrt{2 w_t} \quad \textit{and} \quad P_{\mathcal{L}} \leq \sqrt{2} \sup_{\boldsymbol{z} \in \mathcal{X} \times \mathcal{Y}} \left\{ \frac{\|\nabla_x \mathcal{L}(\boldsymbol{z})\|_{\mathcal{X}^*}}{\sqrt{\mu_{\mathcal{X}}}}, \frac{\|\nabla_y \mathcal{L}(\boldsymbol{z})\|_{\mathcal{Y}^*}}{\sqrt{\mu_{\mathcal{Y}}}} \right\}. \tag{83}$$

Proof. We will first work on $h_t^{(x)}$:

$$\begin{aligned} h_t^{(x)} &= \mathcal{L}(\boldsymbol{x}^{(t)}, \widehat{\boldsymbol{y}}^{(t)}) - \mathcal{L}^* \\ &\leq \mathcal{L}(\boldsymbol{x}^{(t)}, \widehat{\boldsymbol{y}}^{(t)}) - \mathcal{L}(\boldsymbol{x}^*, \widehat{\boldsymbol{y}}^{(t)}) \\ &\leq \left\langle \boldsymbol{x}^{(t)} - \boldsymbol{x}^*, \nabla_x \mathcal{L}(\boldsymbol{x}^{(t)}, \widehat{\boldsymbol{y}}^{(t)}) \right\rangle \qquad \text{(by convexity)} \\ &\leq P_{\mathcal{X}} \sqrt{w_t^{(x)}} \qquad \text{(def of } P_{\mathcal{X}} \text{ (78))}. \end{aligned}$$

We can do the same thing for $h_t^{(y)}$ and $w_t^{(y)}$, thus

$$h_t \leq P_{\mathcal{L}} \left(\sqrt{w_t^{(x)}} + \sqrt{w_t^{(y)}} \right) \leq P_{\mathcal{L}} \sqrt{2 w_t}, \tag{84}$$

where the last inequality uses $\sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}$. Finally, the inequality on $P_{\mathcal{L}}$ is from Proposition 14. □

C Relations between primal suboptimalities and dual gaps

C.1 Primal suboptimalities

Recall that we introduced $\widehat{\boldsymbol{x}}^{(t)} := \arg\min_{\boldsymbol{x} \in \mathcal{X}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^{(t)})$ and similarly $\widehat{\boldsymbol{y}}^{(t)} := \arg\max_{\boldsymbol{y} \in \mathcal{Y}} \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y})$. Then the primal suboptimality is the positive quantity

$$h_t := \mathcal{L}(\boldsymbol{x}^{(t)}, \widehat{\boldsymbol{y}}^{(t)}) - \mathcal{L}(\widehat{\boldsymbol{x}}^{(t)}, \boldsymbol{y}^{(t)}). \tag{85}$$

To get a convergence rate, one has to upper bound the primal suboptimality defined in (85), but it is hard to work with the moving quantities $\widehat{\boldsymbol{x}}^{(t)}$ and $\widehat{\boldsymbol{y}}^{(t)}$ in the analysis. This is why we use in our analysis a different merit function that uses the (fixed) saddle point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ of $\mathcal{L}$ in its definition. We recall its definition below.

Second primal suboptimality. We define the second primal suboptimality for $\mathcal{L}$ of the iterate $(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})$ with respect to the saddle point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ as the positive quantity:

$$w_t := \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}^{(t)}). \tag{86}$$

It follows from $\mathcal{L}(\boldsymbol{x}^{(t)}, \widehat{\boldsymbol{y}}^{(t)}) \geq \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*)$ and $\mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}^{(t)}) \geq \mathcal{L}(\widehat{\boldsymbol{x}}^{(t)}, \boldsymbol{y}^{(t)})$ that $w_t \leq h_t$. Furthermore, under the assumption of uniform strong convex-concavity, we proved in Proposition 15 that the square root of w_t upper bounds h_t up to a constant.

C.2 Gap inequalities

In this section, we will prove the crucial inequalities relating suboptimalities and the gap function. Let's recall the definition of $\boldsymbol{s}^{(t)}$ and $\boldsymbol{v}^{(t)}$:

$$\boldsymbol{s}^{(t)} := \underset{\boldsymbol{s}\in\mathcal{X}\times\mathcal{Y}}{\arg\min} \left\langle \boldsymbol{s}, \boldsymbol{r}^{(t)} \right\rangle \quad \text{and} \quad \boldsymbol{v}^{(t)} := \underset{\boldsymbol{v}\in\mathcal{S}_x^t\times\mathcal{S}_y^t}{\arg\max} \left\langle \boldsymbol{v}, \boldsymbol{r}^{(t)} \right\rangle \tag{87}$$

where $(\boldsymbol{r}^{(t)})^\top := ((\boldsymbol{r}_x^{(t)})^\top, (\boldsymbol{r}_y^{(t)})^\top) := \big(\nabla_x \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}), -\nabla_y \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})\big)$. Also, the following various gaps are defined as

$$g_t^{\text{FW}} := \left\langle \boldsymbol{d}_{\text{FW}}^{(t)}, -\boldsymbol{r}^{(t)} \right\rangle, \qquad g_t^{\text{PFW}} := \left\langle \boldsymbol{d}_{\text{PFW}}^{(t)}, -\boldsymbol{r}^{(t)} \right\rangle \quad \text{and} \quad g_t := \left\langle \boldsymbol{d}^{(t)}, -\boldsymbol{r}^{(t)} \right\rangle \tag{88}$$

where $\boldsymbol{d}_{\text{FW}}^{(t)} = \boldsymbol{s}^{(t)} - \boldsymbol{z}^{(t)}$ and $\boldsymbol{d}_{\text{PFW}}^{(t)} = \boldsymbol{s}^{(t)} - \boldsymbol{v}^{(t)}$. The direction $\boldsymbol{d}^{(t)}$ is the direction chosen by the algorithm at step t: it is always $\boldsymbol{d}_{FW}^{(t)}$ for SP-FW, and can be either $\boldsymbol{d}_{FW}^{(t)}$ or $\boldsymbol{d}_{\text{A}}^{(t)} := \boldsymbol{z}^{(t)} - \boldsymbol{v}^{(t)}$ for SP-AFW. Even if the definitions of these gaps are different, the formalism for the analysis of the convergence of both algorithms is going to be fairly similar. It is straightforward to notice that $g_t^{\text{PFW}} \geq g_t$ and one can show that the current gap g_t is lower bounded by half of g_t^{PFW}:

Lemma 16. *For the SP-AFW algorithm, the current gap g_t can be bounded as follows:*

$$\frac{1}{2} g_t^{\text{PFW}} \leq g_t \leq g_t^{\text{PFW}} \tag{89}$$

Proof. First let's show the RHS of the inequality,

$$g_t^{\text{PFW}} := \left\langle \boldsymbol{d}_{\text{PFW}}^{(t)}, -\boldsymbol{r}^{(t)} \right\rangle = \left\langle \boldsymbol{d}_{\text{A}}^{(t)}, -\boldsymbol{r}^{(t)} \right\rangle + \left\langle \boldsymbol{d}_{\text{FW}}^{(t)}, -\boldsymbol{r}^{(t)} \right\rangle \geq \left\langle \boldsymbol{d}^{(t)}, -\boldsymbol{r}^{(t)} \right\rangle \tag{90}$$

because both $\left\langle \boldsymbol{d}_{\text{A}}^{(t)}, -\boldsymbol{r}^{(t)} \right\rangle \geq 0$ and $\left\langle \boldsymbol{d}_{\text{FW}}^{(t)}, -\boldsymbol{r}^{(t)} \right\rangle \geq 0$ from their definition. For the LHS inequality, we use the fact that $g_t = \max\left\{ \left\langle \boldsymbol{d}_{\text{A}}^{(t)}, -\boldsymbol{r}^{(t)} \right\rangle, \left\langle \boldsymbol{d}_{\text{FW}}^{(t)}, -\boldsymbol{r}^{(t)} \right\rangle \right\}$ for SP-AFW and thus:

$$g_t^{\text{PFW}} = \left\langle \boldsymbol{d}_{\text{A}}^{(t)}, -\boldsymbol{r}^{(t)} \right\rangle + \left\langle \boldsymbol{d}_{\text{FW}}^{(t)}, -\boldsymbol{r}^{(t)} \right\rangle \leq 2 g_t. \tag{91}$$

□

In the following, we will assume that we are in one of the two following cases:

(I) The saddle point of $\mathcal{L}$ belongs to the relative interior of $\mathcal{X} \times \mathcal{Y}$.

(P) $\mathcal{X}$ and $\mathcal{Y}$ are polytopes, i.e. $\exists \mathcal{A}, \mathcal{B}$ finite s.t $\mathcal{X} = \text{conv}(\mathcal{A})$, $\mathcal{Y} = \text{conv}(\mathcal{B})$.

Then either $\mu_{\mathcal{L}}^{\text{int}} > 0$ (case I) or $\mu_{\mathcal{L}}^{\text{A}} > 0$ (case P). Let's write the gap function as the sum of two smaller gap functions:

$$g_t = \underbrace{\left\langle \boldsymbol{d}_{(x)}^{(t)}, -\boldsymbol{r}_x^{(t)} \right\rangle}_{=: g_t^{(x)}} + \underbrace{\left\langle \boldsymbol{d}_{(y)}^{(t)}, -\boldsymbol{r}_y^{(t)} \right\rangle}_{=: g_t^{(y)}} \tag{92}$$

Because of the convex-concavity of $\mathcal{L}$, this scalar product bounds the differences between the value of $\mathcal{L}$ at the point $(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})$ and the value of $\mathcal{L}$ at another point. Hence this gap function upper-bounds h_t and w_t defined in (85) and (86). More concretely, we have the following lemma.

Lemma 17. *For all t in $\mathbb{N}$, $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}$*

$$g_t^{\mathrm{PFW}} \geq g_t^{\mathrm{FW}} \geq \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^{(t)}), \tag{93}$$

and, furthermore,

$$g_t \geq h_t \geq w_t. \tag{94}$$

Proof. First let's show the LHS of (93),

$$g_t^{\mathrm{PFW}} = \left\langle \boldsymbol{d}_{\mathrm{PFW}}^{(t)}, -\boldsymbol{r}^{(t)} \right\rangle = \left\langle \boldsymbol{d}_{\mathrm{A}}^{(t)}, -\boldsymbol{r}^{(t)} \right\rangle + \left\langle \boldsymbol{d}_{\mathrm{FW}}^{(t)}, -\boldsymbol{r}^{(t)} \right\rangle \geq \left\langle \boldsymbol{d}_{\mathrm{FW}}^{(t)}, -\boldsymbol{r}^{(t)} \right\rangle = g_t^{\mathrm{FW}} \tag{95}$$

because one can easily derive that $\left\langle \boldsymbol{d}_{\mathrm{A}}^{(t)}, -\boldsymbol{r}^{(t)} \right\rangle \geq 0$ from the definition of the away direction $\boldsymbol{d}_{\mathrm{A}}^{(t)}$. It follows from convexity of $\boldsymbol{x} \mapsto \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^{(t)})$ that for all $\boldsymbol{x}$ in $\mathcal{X}$,

$$(g_t^{\mathrm{FW}})_x := \left\langle (\boldsymbol{d}_{\mathrm{FW}}^{(t)})_x, -\nabla_x \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \right\rangle \geq \left\langle \boldsymbol{x} - \boldsymbol{x}^{(t)}, -\nabla_x \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \right\rangle \tag{96}$$

$$\geq \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^{(t)}). \tag{97}$$

A similar inequality emerges through the convexity of $\boldsymbol{y} \mapsto -\mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y})$,

$$(g_t^{\mathrm{FW}})_y := \left\langle (\boldsymbol{d}_{\mathrm{FW}}^{(t)})_y, \nabla_y \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \right\rangle \geq \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}) - \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}), \tag{98}$$

which gives us

$$g_t^{\mathrm{FW}} \geq \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^{(t)}) + \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}) - \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}), \tag{99}$$

which shows (93). By using $\boldsymbol{x} = \widehat{\boldsymbol{x}}^{(t)}$ and $\boldsymbol{y} = \widehat{\boldsymbol{y}}^{(t)}$ in (93), we get $g_t^{\mathrm{FW}} \geq h_t$. We also know that $g_t = \max(g_t^{\mathrm{A}}, g_t^{\mathrm{FW}}) \geq g_t^{\mathrm{FW}}$ for SP-AFW. So combining with $h_t \geq w_t$ that we already knew, we get (94). □

Next, we recall two lemmas, one from Lacoste-Julien and Jaggi (2013) and the other one from (Lacoste-Julien and Jaggi, 2015). These lemmas upper bound the primal suboptimality with the square of the gap times a constant depending on the geometric (or the interior) strong convexity constant.

Lemma 18 (Lacoste-Julien and Jaggi (2015), Lacoste-Julien and Jaggi (2013))**.** *If f is strongly convex, then for any $\boldsymbol{x}^{(t)} \in \mathcal{X}$,*

$$f(\boldsymbol{x}^{(t)}) - f(\boldsymbol{x}_c) \leq \frac{\left(g_t^{\mathrm{FW}}\right)^2}{2\mu_f^{\boldsymbol{x}_c}} \quad \text{if } \boldsymbol{x}_c \in \text{ interior of } \mathcal{X} \quad \textit{(Lacoste-Julien and Jaggi, 2013)} \tag{100}$$

and

$$f(\boldsymbol{x}^{(t)}) - f^* \leq \frac{\left(g_t^{\mathrm{PFW}}\right)^2}{2\mu_f^{\mathrm{A}}} \quad \text{if } \mathcal{X} = \operatorname{conv}(\mathcal{A}) \quad \textit{(Lacoste-Julien and Jaggi, 2015)} \tag{101}$$

where $g_t^{\mathrm{FW}} = \langle \boldsymbol{x}^{(t)} - \boldsymbol{s}^{(t)}, \nabla f(\boldsymbol{x}^{(t)}) \rangle$, $g_t^{\mathrm{PFW}} = \langle \boldsymbol{v}^{(t)} - \boldsymbol{s}^{(t)}, \nabla f(\boldsymbol{x}^{(t)}) \rangle$ and $f^* = \min_{x \in \mathcal{X}} f$.

Notice once again that in this lemma *we do not need $\boldsymbol{x}_c$ to be optimal.*

Proof. Let $\boldsymbol{x}^{(t)} \neq \boldsymbol{x}_c$. Using the definition of interior strong convexity (46) and choosing γ such that $\boldsymbol{x}_c = \boldsymbol{x}^{(t)} + \gamma \left(\bar{\boldsymbol{s}} \left(\boldsymbol{x}^{(t)}, \boldsymbol{x}_c \right) - \boldsymbol{x}^{(t)} \right)$, we get

$$\begin{aligned} f(\boldsymbol{x}_c) - f(\boldsymbol{x}^{(t)}) &\geq \gamma \left\langle \bar{\boldsymbol{s}}(\boldsymbol{x}_c, \boldsymbol{x}^{(t)}) - \boldsymbol{x}^{(t)}, \nabla f(\boldsymbol{x}^{(t)}) \right\rangle + \gamma^2 \frac{\mu_f^{\boldsymbol{x}_c}}{2} \\ &\geq -\gamma g_t^{\mathrm{FW}} + \gamma^2 \frac{\mu_f^{\boldsymbol{x}_c}}{2} \\ &\geq \frac{-\left(g_t^{\mathrm{FW}}\right)^2}{2\mu_f^{\boldsymbol{x}_c}}. \end{aligned}$$

The last line of this derivation is obtained through the inequality: $-a^2 + 2ab - b^2 \leq 0$. If $\boldsymbol{x}^{(t)} = \boldsymbol{x}_c$ the inequality is just the positivity of the gap.

For the second statement, we will use the definition of the geometric strong convexity constant (Equation (53)) at the point $\boldsymbol{x} = \boldsymbol{x}^{(t)}$ and $\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} f(x)$. Recall that $\gamma^{\mathrm{A}}(\boldsymbol{x}, \boldsymbol{x}^*) = \frac{\langle -\nabla f(\boldsymbol{x}), \boldsymbol{x}^* - \boldsymbol{x}\rangle}{\langle -\nabla f(\boldsymbol{x}), \boldsymbol{s}_f(\boldsymbol{x}) - \boldsymbol{v}_f(\boldsymbol{x})\rangle}$.

$$
\begin{aligned}
f(\boldsymbol{x}^*) - f(\boldsymbol{x}^{(t)}) &\geq \left\langle \boldsymbol{x}^* - \boldsymbol{x}^{(t)}, \nabla f(\boldsymbol{x}^{(t)}) \right\rangle + \frac{\mu_f^{\mathrm{A}}}{2}\gamma^{\mathrm{A}}(\boldsymbol{x}^{(t)}, \boldsymbol{x}^*)^2 \\
&= -\gamma^{\mathrm{A}}(\boldsymbol{x}^{(t)}, \boldsymbol{x}^*)\left\langle \boldsymbol{s}_f(\boldsymbol{x}^{(t)}) - \boldsymbol{v}_f(\boldsymbol{x}^{(t)}), -\nabla f(\boldsymbol{x}^{(t)}) \right\rangle + \frac{\mu_f^{\mathrm{A}}}{2}\gamma^{\mathrm{A}}(\boldsymbol{x}^{(t)}, \boldsymbol{x}^*)^2 \\
&\geq -\gamma^{\mathrm{A}}(\boldsymbol{x}^{(t)}, \boldsymbol{x}^*) g_t^{\mathrm{PFW}} + \frac{\mu_f^{\mathrm{A}}}{2}\gamma^{\mathrm{A}}(\boldsymbol{x}^{(t)}, \boldsymbol{x}^*)^2 \qquad \text{(Equation (54))} \\
&\geq \frac{-\left(g_t^{\mathrm{PFW}}\right)^2}{2\mu_f^{\mathrm{A}}}.
\end{aligned}
$$

□

Lemma 18 is useful to understand the following lemma and its proof which is just an extension to the convex-concave case.

Lemma 19 (Quadratic gap upper bound on second suboptimality for (I) or (P))**.** *If $\mathcal{L}$ is a strongly convex-concave function, then for any $(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \in \mathcal{X} \times \mathcal{Y}$,*

$$
w_t \leq \frac{(g_t^{\mathrm{FW}})^2}{2\mu_{\mathcal{L}}^{int}} \quad \textit{for (I)} \quad \textit{and} \quad w_t \leq h_t \leq \frac{(g_t^{\mathrm{PFW}})^2}{2\mu_{\mathcal{L}}^{\mathrm{A}}} \quad \textit{for (P)} \tag{102}
$$

where the gaps are defined in (88), $\mu_{\mathcal{L}}^{int} := \min\{\mu_{\mathcal{L}}^{\boldsymbol{x}^*}, \mu_{\mathcal{L}}^{\boldsymbol{y}^*}\}$ (i.e. using the reference points $(\boldsymbol{x}_c, \boldsymbol{y}_c) := (\boldsymbol{x}^*, \boldsymbol{y}^*)$ in the definition (66)) and $\mu_{\mathcal{L}}^{\mathrm{A}}$ is the geometric strong convex-concavity of $\mathcal{L}$ over $\mathcal{A} \times \mathcal{B}$, as defined in (71).

Proof. For (I):

Let the function f on $\mathcal{X}$ be defined by $f(\boldsymbol{x}) = \mathcal{L}(\boldsymbol{x}', \boldsymbol{y}^{(t)})$, and the function g on $\mathcal{Y}$ be $g(\boldsymbol{y}') = -\mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}')$. Then using the Lemma 18 on the function f with the reference point $\boldsymbol{x}^*$, and on g with reference point $\boldsymbol{y}^*$, we get

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) - \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}^{(t)}) &\leq \frac{\left\langle \boldsymbol{s}_x^{(t)} - \boldsymbol{x}^{(t)}, -\nabla_x \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \right\rangle^2}{2\mu_f^{\boldsymbol{x}^*}} \\
\mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) &\leq \frac{\left\langle \boldsymbol{s}_y^{(t)} - \boldsymbol{y}^{(t)}, \nabla_y \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \right\rangle^2}{2\mu_g^{\boldsymbol{y}^*}}.
\end{aligned}
$$

As $\mu_{\mathcal{L}}^{\mathrm{int}}$ is smaller than both $\mu_f^{\boldsymbol{x}^*}$ and $\mu_g^{\boldsymbol{y}^*}$ by the definition (66), we can use it in the denominator of the above two inequalities. As we saw from Section C.2 in (92), the gap can be split as sum of the gap of the block $\mathcal{X}$ and the gap of the block $\mathcal{Y}$, i.e. $g_t^{\mathrm{FW}} = \left\langle \boldsymbol{s}_x^{(t)} - \boldsymbol{x}^{(t)}, -\nabla_x \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \right\rangle + \left\langle \boldsymbol{s}_y^{(t)} - \boldsymbol{y}^{(t)}, \nabla_y \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \right\rangle$. Then, using the inequality: $a^2 + b^2 \leq (a+b)^2$ for $(a, b \geq 0)$, we obtain

$$
w_t \leq \frac{(g_t^{\mathrm{FW}})^2}{2\mu_{\mathcal{L}}^{\mathrm{int}}}. \tag{103}
$$

For (P):

Using the Lemma 18 for case (P) on the same functions f and g defined above, we get

$$\mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) - \mathcal{L}(\widehat{\boldsymbol{x}}^{(t)}, \boldsymbol{y}^{(t)}) \leq \frac{\left\langle \boldsymbol{s}_x^{(t)} - \boldsymbol{v}_x^{(t)}, \nabla_x \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \right\rangle^2}{2\mu_f^{\mathrm{A}}}$$

$$\mathcal{L}(\boldsymbol{x}^{(t)}, \widehat{\boldsymbol{y}}^{(t)}) - \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \leq \frac{\left\langle \boldsymbol{s}_y^{(t)} - \boldsymbol{v}_y^{(t)}, -\nabla_y \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \right\rangle^2}{2\mu_g^{\mathrm{A}}}.$$

Using a similar argument as the one to get (103), using that $\mu_{\mathcal{L}}^{\mathrm{A}}$ is smaller than both μ_f^{A} and μ_g^{A}, and referring to the separation of the gap (92), we get

$$h_t \leq \frac{(g_t^{\mathrm{PFW}})^2}{2\mu_{\mathcal{L}}^{\mathrm{A}}}. \tag{104}$$

□

D Convergence analysis

In this section, we are going to show two important lemmas. The first one shows that under some assumptions we can get a Frank-Wolfe-style induction scheme relating the second suboptimality of the potential update w_γ, the current value of the second suboptimality w_t, the gap g_t and any step size $\gamma \in [0, \gamma_{\max}]$. The second lemma will relate the gap and the square root of w_t; this relation enables us to get a rate on the gap after getting a rate on w_t.

D.1 First lemmas

The first lemma in this section is inspired from the standard FW progress lemma, such as Lemma C.2 in (Lacoste-Julien et al., 2013), though it requires a non-trivial change due to the compensation phenomenon for $\mathcal{L}$ mentioned in the main text in (10). In the following, we define the possible updated iterate $\boldsymbol{z}_\gamma$ for $\gamma \in [0, \gamma_{\max}]$:

$$\boldsymbol{z}_\gamma := (\boldsymbol{x}_\gamma, \boldsymbol{y}_\gamma) := \boldsymbol{z}^{(t)} + \gamma \boldsymbol{d}^{(t)}, \quad \text{where} \quad \boldsymbol{d}^{(t)} \text{ is the direction of the step.} \tag{105}$$

For a FW step $\boldsymbol{d}^{(t)} = \boldsymbol{d}_{\mathrm{FW}}^{(t)} := \boldsymbol{s}^{(t)} - \boldsymbol{z}^{(t)}$ and for an away step $\boldsymbol{d}^{(t)} = \boldsymbol{d}_{\mathrm{A}}^{(t)} := \boldsymbol{z}^{(t)} - \boldsymbol{v}^{(t)}$. We also define the corresponding new suboptimality for $\boldsymbol{z}_\gamma$:

$$w_\gamma := \mathcal{L}(\boldsymbol{x}_\gamma, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}_\gamma). \tag{106}$$

Lemma 20 (Suboptimality progress for SP-FW and SP-AFW)**.** *Let $\mathcal{L}$ be strongly convex-concave,*

If we are in case (I) and $\boldsymbol{d}^{(t)} = \boldsymbol{d}_{\mathrm{FW}}^{(t)}$ is a FW direction, we have for any $\gamma \in [0, 1]$:

$$w_\gamma \leq w_t - \nu^{\mathrm{FW}} \gamma g_t^{\mathrm{FW}} + \gamma^2 C_{\mathcal{L}}, \quad \textit{where} \quad \nu^{\mathrm{FW}} := 1 - \frac{M_{\mathcal{L}}}{\sqrt{\mu_{\mathcal{L}}^{int}}}. \tag{107}$$

If we are in case (P) and $\boldsymbol{d}^{(t)}$ is defined from a step of SP-AFW (Algorithm 3), we have for any $\gamma \in [0, \gamma_{\max}]$:

$$w_\gamma \leq w_t - \nu^{\mathrm{PFW}} \gamma g_t^{\mathrm{PFW}} + \gamma^2 C_{\mathcal{L}}, \quad \textit{where} \quad \nu^{\mathrm{PFW}} := \frac{1}{2} - \frac{M_{\mathcal{L}}}{\sqrt{\mu_{\mathcal{L}}^{\mathrm{A}}}}. \tag{108}$$

Proof. The beginning of the argument works with any direction $\boldsymbol{d}^{(t)}$. Recall that $w_t^{(x)} := \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*) - \mathcal{L}^*$ and $w_t^{(y)} := -\mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}^{(t)}) + \mathcal{L}^*$. Now writing $\boldsymbol{x}_\gamma = \boldsymbol{x}^{(t)} + \gamma \boldsymbol{d}_x^{(t)}$ and using the definition of the curvature C_f (41) for the function $\boldsymbol{x} \mapsto f(\boldsymbol{x}) := \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^*)$, we get

$$\begin{aligned} w_\gamma^{(x)} &:= \mathcal{L}(\boldsymbol{x}_\gamma, \boldsymbol{y}^*) - \mathcal{L}^* \\ &\leq \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*) - \mathcal{L}^* + \gamma \left\langle \boldsymbol{d}_x^{(t)}, \nabla_x \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*) \right\rangle + \gamma^2 \frac{C_\mathcal{L}}{2}, \end{aligned} \tag{109}$$

since $C_\mathcal{L} \leq C_f$ by definition (58). Recall that the gap function g_t can be decomposed by (92) into two smaller gap functions $g_t^{(x)} := \left\langle \boldsymbol{d}_x^{(t)}, -\boldsymbol{r}_x^{(t)} \right\rangle$ and $g_t^{(y)} := \left\langle \boldsymbol{d}_y^{(t)}, -\boldsymbol{r}_y^{(t)} \right\rangle$. We define $\epsilon_t := \left\langle \boldsymbol{d}_x^{(t)}, \nabla_x \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*) - \nabla_x \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \right\rangle$ to be the sequence representing the error between the gradient used for the minimization and the gradient at the point $(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*)$. Then,

$$w_\gamma^{(x)} \leq w_t^{(x)} - \gamma g_t^{(x)} + \gamma \epsilon_t + \gamma^2 \frac{C_\mathcal{L}}{2}. \tag{110}$$

Now, as M_{XY} is finite (under Lipschitz gradient assumption), we can use the definition of the bilinearity constant (73) to get

$$|\epsilon_t| = \left| \left\langle \boldsymbol{d}_x^{(t)}, \nabla_x \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*) - \nabla_x \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \right\rangle \right| \leq \sqrt{w_t^{(y)}} M_{XY}. \tag{111}$$

Combining equations (110) and (111) we finally obtain

$$w_\gamma^{(x)} \leq w_t^{(x)} - \gamma g_t^{(x)} + \gamma M_{XY} \sqrt{w_t^{(y)}} + \gamma^2 \frac{C_\mathcal{L}}{2}. \tag{112}$$

We can get an analogous inequality for $w_\gamma^{(y)}$,

$$w_\gamma^{(y)} \leq w_t^{(y)} - \gamma g_t^{(y)} + \gamma M_{YX} \sqrt{w_t^{(x)}} + \gamma^2 \frac{C_\mathcal{L}}{2}. \tag{113}$$

Then adding $w_\gamma^{(x)}$ and $w_\gamma^{(y)}$ and using $\sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}$ (coming from the concavity of $\sqrt{\cdot}$), we get

$$w_\gamma \leq w_t - \gamma g_t + \gamma M_\mathcal{L} \sqrt{2 w_t} + 2\gamma^2 \frac{C_\mathcal{L}}{2}. \tag{114}$$

We stress that the above inequality (114) is valid for any direction $\boldsymbol{d}^{(t)}$, using $g_t := \left\langle \boldsymbol{d}^{(t)}, -\boldsymbol{r}^{(t)} \right\rangle$, and for any feasible step size γ such that $\boldsymbol{z}_\gamma \in \mathcal{X} \times \mathcal{Y}$ (the last condition was used in the definition of C_f; see also footnote 2 for more information).

To finish the argument, we now use the specific property of the direction $\boldsymbol{d}^{(t)}$ and use the crucial Lemma 19 that relates w_t with the square of the appropriate gap.

For the case (I) of interior saddle point, we consider $\boldsymbol{d}^{(t)} = \boldsymbol{d}_{\mathrm{FW}}^{(t)}$ and thus $g_t = g_t^{\mathrm{FW}}$. Then combining Lemma 19 (using the interior strong convexity constant) with (114), we get

$$w_\gamma \leq w_t - \gamma \left(1 - \frac{M_\mathcal{L}}{\sqrt{\mu_\mathcal{L}^{\mathrm{int}}}} \right) g_t^{\mathrm{FW}} + \gamma^2 C_\mathcal{L}. \tag{115}$$

For the case (P) of polytope domains, we consider $\boldsymbol{d}^{(t)}$ as defined by the SP-AFW algorithm. We thus have $g_t \geq \frac{1}{2} g_t^{\mathrm{PFW}}$ by Lemma 16. Then combining Lemma 19 (using the geometric strong convexity constant) with (114), we get

$$w_\gamma \leq w_t - \gamma \left(\frac{1}{2} - \frac{M_\mathcal{L}}{\sqrt{\mu_\mathcal{L}^{\mathrm{A}}}} \right) g_t^{\mathrm{PFW}} + \gamma^2 C_\mathcal{L}, \tag{116}$$

which finishes the proof. Still in the case (P), we also present an inequality in terms of the direction gap g_t (which yields a better constant that will be important for the sublinear convergence proof in Theorem 25) by using instead the inequality $g_t^{\text{PFW}} \geq g_t$ (Lemma 16) with Lemma 19 and (114):

$$w_\gamma \leq w_t - \gamma \left(1 - \frac{M_{\mathcal{L}}}{\sqrt{\mu_{\mathcal{L}}^{\text{A}}}}\right) g_t + \gamma^2 C_{\mathcal{L}}. \tag{117}$$

□

The above lemma uses a specific update direction $\boldsymbol{d}^{(t)}$ to get a potential new suboptimality w_γ. By using the property that $w_\gamma \geq 0$ always, we can actually derive an upper bound on the gap in terms of w_t *irrespective of any algorithm* (i.e. this relationship holds for any possible feasible point $(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})$). More precisely, for SP-FW algorithm (case (I)) the only thing we need to set is a feasible point $\boldsymbol{z}^{(t)}$ but for the SP-AFW algorithm (case (P)) we also need an active set expansion for $\boldsymbol{z}^{(t)}$ for which the maximum away step size is larger than $\frac{\nu g_t}{2C_{\mathcal{L}}}$ (which can potentially not be the active set calculated by an algorithm). This is stated in the following theorem, which is a saddle point generalization of the gap upper bound given in Theorem 2 of (Lacoste-Julien and Jaggi, 2015).

Theorem 21 (Bounding the gap with the second suboptimality)**.** *If $\mathcal{L}$ is strongly convex-concave and has a finite curvature constant then*

- *case (I): For any $\boldsymbol{z}^{(t)} \in \mathcal{X} \times \mathcal{Y}$,*

$$g_t^{\text{FW}} \leq \frac{2}{\nu^{\text{FW}}} \max\left\{\sqrt{C_{\mathcal{L}} w_t}, w_t\right\}. \tag{118}$$

  Since $\boldsymbol{z}^{(t)}$ is fixed, this statement is algorithm free.

- *case (P): For any $\boldsymbol{z}^{(t)} \in \mathcal{X} \times \mathcal{Y}$, if there exists an active set expansion for $\boldsymbol{z}^{(t)}$ for which $\gamma_{\max} = 1$ or $\gamma_{\max} \geq \frac{\nu^{\text{PFW}} g_t^{\text{PFW}}}{2C_{\mathcal{L}}}$ (see (122) for the definition of $\gamma_{\max}$) then,*

$$g_t^{\text{PFW}} \leq \frac{2}{\nu^{\text{PFW}}} \max\left\{\sqrt{C_{\mathcal{L}} w_t}, w_t\right\}. \tag{119}$$

Both statement are algorithm free but g_t^{PFW} depends on a chosen expansion of $\boldsymbol{z}^{(t)}$:

$$\boldsymbol{z}^{(t)} = \sum_{(\boldsymbol{v}_x, \boldsymbol{v}_y) \in \mathcal{S}^{(t)}} \alpha_{\boldsymbol{v}_x}^{(t)} \alpha_{\boldsymbol{v}_y}^{(t)} (\boldsymbol{v}_x, \boldsymbol{v}_y) \quad \textit{where} \quad \mathcal{S}^{(t)} := \left\{(\boldsymbol{v}_x, \boldsymbol{v}_y) \in \mathcal{A} \times \mathcal{B} \,;\, \alpha_{\boldsymbol{v}_x}^{(t)} \alpha_{\boldsymbol{v}_y}^{(t)} > 0\right\}, \tag{120}$$

because,

$$\begin{aligned} g_t^{\text{PFW}} &:= \langle -\boldsymbol{r}^{(t)}, \boldsymbol{d}_{\text{FW}}^{(t)} + \boldsymbol{d}_{\text{A}}^{(t)} \rangle \\ \textit{where} \quad \boldsymbol{d}_{\text{A}}^{(t)} &:= \boldsymbol{z}^{(t)} - \arg\max_{\boldsymbol{v} \in \mathcal{S}_x \times \mathcal{S}_y} \langle \boldsymbol{r}^{(t)}, \boldsymbol{v} \rangle, \\ \textit{and} \quad \boldsymbol{d}_{\text{FW}}^{(t)} &:= \arg\min_{\boldsymbol{v} \in \mathcal{A} \times \mathcal{B}} \langle \boldsymbol{r}^{(t)}, \boldsymbol{v} \rangle - \boldsymbol{z}^{(t)}. \end{aligned} \tag{121}$$

The maximum step size associated with the active set expansion described in Equation (120) is

$$\gamma_{\max} := \begin{cases} 1 \quad \textit{if} \quad \left\langle -\boldsymbol{r}^{(t)}, \boldsymbol{d}_{\text{A}}^{(t)} \right\rangle \leq \left\langle -\boldsymbol{r}^{(t)}, \boldsymbol{d}_{\text{FW}}^{(t)} \right\rangle, \\ \min\left\{ \frac{\alpha_{\boldsymbol{v}_x^{(t)}}^{(t)}}{1 - \alpha_{\boldsymbol{v}_x^{(t)}}^{(t)}}, \frac{\alpha_{\boldsymbol{v}_y^{(t)}}^{(t)}}{1 - \alpha_{\boldsymbol{v}_y^{(t)}}^{(t)}} \right\} \quad \textit{otherwise.} \end{cases} \tag{122}$$

Proof. In this proof, we let (g_t, ν) to stand respectively for $(g_t^{\mathrm{FW}}, \nu^{\mathrm{FW}})$ for case (I) or $(g_t^{\mathrm{PFW}}, \nu^{\mathrm{PFW}})$ for case (P). We will start from the inequalities (107) and (108) in Lemma 20. Equation (107) is valid considering a FW direction $\boldsymbol{d}_{\mathrm{FW}}^{(t)}$, Equation (108) is valid if we consider the direction that would have be set by the SP-AFW algorithm if it was run at point $\boldsymbol{z}^{(t)}$ with the active set expansion described in the theorem statement. Since $w_\gamma \geq 0$, for both cases become :

$$0 \leq w_t - \gamma\nu g_t + \gamma^2 C_{\mathcal{L}}, \tag{123}$$

then we can put the gap on the LHS,

$$\gamma\nu g_t - \gamma^2 C_{\mathcal{L}} \leq w_t. \tag{124}$$

This inequality is valid for any $\gamma \in [0, \gamma_{\max}]$. In order to get the tightest bound between the gap and the suboptimality, we will maximize the LHS. It can be maximized with $\bar{\gamma} := \frac{\nu g_t}{2C_{\mathcal{L}}} = \gamma_t$. Now we have two cases:

If $\bar{\gamma} \in [0, \gamma_{\max}]$, then we get: $\nu g_t \leq 2\sqrt{C_{\mathcal{L}}\, w_t}$.

And if $\gamma_{\max} = 1$ and $\bar{\gamma} = \frac{\nu g_t}{2C_{\mathcal{L}}} > 1$, then setting $\gamma = 1$ we get: $\nu g_t \leq 2w_t$. By taking the maximum between the two options, we get the theorem statement. □

The previous theorem guarantees that the gap gets small when w_t gets small (only for the non-drop steps if in situation (P)). As we use the gap as a stopping criterion in the algorithm, this is a useful theorem to provide an upper bound on the number of iterations needed to get a certificate of suboptimality.

The following corollary provides a better bound on h_t than the inequality $h_t \leq \mathrm{cst}\sqrt{w_t}$ previously shown (14) when the situation is (P). It will be useful later to get a better rate of convergence for h_t under hypothesis (P).

Corollary 22 (Tighter bound on h_t for non-drop steps in situation (P))**.** *Suppose that $\mathcal{L}$ is strongly convex-concave and has a finite curvature constant, and that the domain is a product of polytopes (i.e. we are in situation (P)). Let $\boldsymbol{z}^{(t)} \in \mathcal{X} \times \mathcal{Y}$ be given. If there exists an active set expansion for $\boldsymbol{z}^{(t)}$ for which the maximum step size is larger than $\frac{\nu g_t}{2C_{\mathcal{L}}}$ (see Theorem (21) for more details) ,*

$$h_t \leq \frac{2\max\{C_{\mathcal{L}}, w_t\}}{\mu_{\mathcal{L}}^{\mathrm{A}}(\nu^{\mathrm{PFW}})^2} w_t, \tag{125}$$

where ν^{PFW} is defined in (108).

Proof. By Lemma 19 in situation (P), we have:

$$h_t \leq \frac{(g_t^{\mathrm{PFW}})^2}{2\mu_{\mathcal{L}}^{\mathrm{A}}} \leq \frac{2\max\{C_{\mathcal{L}}, w_t\}}{\mu_{\mathcal{L}}^{\mathrm{A}}(\nu^{\mathrm{PFW}})^2} w_t.$$

The last inequality is obtained by applying the upper bound on the gap given in the previous Theorem 21. □

D.2 Proof of Theorem 1

In this section, we will prove that under some conditions on the constant defined in subsections B.2 and B.3, the suboptimalities w_t vanish linearly with the adaptive step size $\gamma_t = \min\left\{\gamma_{\max}, \frac{\nu}{2C_{\mathcal{L}}} g_t\right\}$ or sublinearly with the universal step size $\gamma_t = \min\left\{\gamma_{\max}, \frac{2}{2+k(t)}\right\}$.

Lemma 23 (Geometric decrease of second suboptimality). *Let $\mathcal{L}$ be a strongly convex-concave function with a smoothness constant $C_{\mathcal{L}}$, a positive interior strong convex-concavity constant $\mu_{\mathcal{L}}^{int}$ (66) or a positive geometric strong convex-concavity $\mu_{\mathcal{L}}^{\mathrm{A}}$ (71). Let us also define the rate multipliers ν as*

$$\nu^{\mathrm{FW}} := 1 - \frac{M_{\mathcal{L}}}{\sqrt{\mu_{\mathcal{L}}^{int}}} \quad \text{and} \quad \nu^{\mathrm{PFW}} := \frac{1}{2} - \frac{M_{\mathcal{L}}}{\sqrt{\mu_{\mathcal{L}}^{\mathrm{A}}}} \quad \text{(see Equation (75) for the definition of } M_{\mathcal{L}}\text{)}. \tag{126}$$

Let the tuple $(g_t, \nu, \mu_{\mathcal{L}})$ refers to either $(g_t^{\mathrm{FW}}, \nu^{\mathrm{FW}}, \mu_{\mathcal{L}}^{int})$ for case (I) where the algorithm is SP-FW , or $(g_t^{\mathrm{PFW}}, \nu^{\mathrm{PFW}}, \mu_{\mathcal{L}}^{\mathrm{A}})$ for case (P) where the algorithm is SP-AFW,

If $\nu > 0$, then at each non-drop step (when $\gamma_t < \gamma_{\max}$ or $\gamma_{\max} \geq 1$), the suboptimality w_t of the algorithm with step size $\gamma_t = \min(\gamma_{\max}, \frac{\nu}{2C_{\mathcal{L}}} g_t)$ decreases geometrically as

$$w_{t+1} \leq (1 - \rho_{\mathcal{L}}) w_t \tag{127}$$

where $\rho_{\mathcal{L}} := \frac{\nu^2}{2} \frac{\mu_{\mathcal{L}}}{C_{\mathcal{L}}}$. Moreover, for case (I) there is no drop step and for case (P) the number of drop step (when $\gamma_t = \gamma_{\max}$) is upper bounded by two third of the number of iteration (see Section A, Equation (38)), while when we have a drop step, we still have:

$$w_{t+1} \leq w_t. \tag{128}$$

Proof. The bulk of the proof is of a similar form for both SP-FW and SP-AFW, and so in the following, we let $(g_t, \mu_{\mathcal{L}}, \nu)$ to stand respectively for $(g_t^{\mathrm{FW}}, \mu_{\mathcal{L}}^{\mathrm{int}}, \nu^{\mathrm{FW}})$ for SP-FW (case (I)) or $(g_t^{\mathrm{PFW}}, \mu_{\mathcal{L}}^{\mathrm{A}}, \nu^{\mathrm{PFW}})$ for SP-AFW (case (P)). As $\gamma_t \leq \gamma_{\max}$, we can apply the important Lemma 20 with $\gamma = \gamma_t$ (the actual step size that was taken in the algorithm) to get:

$$w_{t+1} = w_{\gamma_t} \leq w_t - \nu \gamma_t g_t + \gamma_t^2 C_{\mathcal{L}}. \tag{129}$$

We note in passing that the adaptive step size rule $\gamma_t = \min(\gamma_{\max}, \frac{\nu}{2C_{\mathcal{L}}} g_t)$ was specifically chosen to minimize the RHS of (129) among the feasible step sizes.

If $\frac{\nu}{2C_{\mathcal{L}}} g_t \leq \gamma_{\max}$, then we have $\gamma_t = \frac{\nu}{2C_{\mathcal{L}}} g_t$ and so (129) becomes:

$$w_{t+1} \leq w_t - \frac{\nu^2}{2C_{\mathcal{L}}} (g_t)^2 + \frac{\nu^2}{4C_{\mathcal{L}}} (g_t)^2 = w_t - \frac{\nu^2}{4C_{\mathcal{L}}} (g_t)^2. \tag{130}$$

Applying the fact that the square of the appropriate gap upper bounds w_t (Lemma 19 with a similar form for both cases (I) and (P)), we directly obtain the claimed geometric decrease

$$w_{t+1} \leq w_t \left(1 - \frac{\nu^2}{2} \frac{\mu_{\mathcal{L}}}{C_{\mathcal{L}}}\right). \tag{131}$$

If $\frac{\nu}{2C_{\mathcal{L}}} g_t > \gamma_{\max}$, then we have $\gamma_t = \gamma_{\max}$ and so (129) becomes:

$$\begin{aligned} w_{t+1} &\leq w_t - \nu \gamma_{\max} g_t + \gamma_{\max}^2 C_{\mathcal{L}} \\ &\leq w_t - \nu \gamma_{\max} g_t + \frac{\nu}{2} \gamma_{\max} g_t && \text{(using} \quad C_{\mathcal{L}} < \frac{\nu}{2\gamma_{\max}} g_t\text{)} \qquad (132) \\ &\leq w_t - \frac{\nu}{2} \gamma_{\max} g_t \\ &\leq w_t \left(1 - \frac{\nu}{2} \gamma_{\max}\right). && (w_t \leq g_t \text{ by Lemma 17}) \qquad (133) \end{aligned}$$

If $\gamma_{\max} \geq 1$ (either we are taking a FW step or an away step with a big step size), then the geometric rate is at least $(1 - \frac{\nu}{2})$, which is a better rate than $\rho_{\mathcal{L}}$ since $\nu^2 \leq \nu$ as $\nu \leq 1$, and one can show that $\frac{\mu_{\mathcal{L}}}{C_{\mathcal{L}}} \leq 1$ always (see Remark 7 in Appendix D of (Lacoste-Julien and Jaggi, 2015) for case (P) and use

a similar argument for case (I)). Thus $\rho_{\mathcal{L}}$ is valid both when $\gamma_t < \gamma_{\max}$ or $\gamma_{\max} \geq 1$, as claimed in the theorem.

When $\gamma_t = \gamma_{\max} < 1$, we cannot guarantee sufficient progress as $\gamma_{\max}$ could be arbitrarily small (this can only happen for an away step as $\gamma_{\max} = 1$ for a FW step). These are the problematic *drop steps*, but as explained in Appendix A with Equation (38), they cannot happen too often for SP-AFW.

Finally, to show that the suboptimality cannot increase during a drop step ($\gamma_t = \gamma_{\max}$), we point out that the function $\gamma \mapsto w_t - \gamma\nu^{\text{PFW}} g_t^{\text{PFW}} + \gamma^2 C_{\mathcal{L}}$ is a convex function that is minimized by $\bar{\gamma} = \frac{\nu^{\text{PFW}}}{2C_{\mathcal{L}}} g_t^{\text{PFW}}$ and so is decreasing on $[0, \bar{\gamma}]$. When $\gamma_t = \gamma_{\max}$, we have that $\gamma_{\max} \leq \bar{\gamma}$, and thus the value for $\gamma = \gamma_{\max}$ is lower than the value for $\gamma = 0$, i.e.

$$w_{t+1} \leq w_t - \gamma_{\max}\nu^{\text{PFW}} g_t^{\text{PFW}} + \gamma_{\max}^2 C_{\mathcal{L}} \leq w_t. \tag{134}$$

□

The previous lemma (Lemma 23), the fact that the gap upper bounds the suboptimality (Lemma 17) and the primal suboptimalities analysis lead us directly to the following theorem. This theorem is the affine invariant formulation with adaptive step size of Theorem 1.

Theorem 24. *Let $\mathcal{L}$ be a strongly convex-concave function with a finite smoothness constant $C_{\mathcal{L}}$, a positive interior strong convex-concavity constant $\mu_{\mathcal{L}}^{int}$ (66) or a positive geometric strong convex-concavity $\mu_{\mathcal{L}}^{\text{A}}$ (71). Let us also define the rate multipliers ν as*

$$\nu^{\text{FW}} := 1 - \frac{M_{\mathcal{L}}}{\sqrt{\mu_{\mathcal{L}}^{int}}} \quad \text{and} \quad \nu^{\text{PFW}} := \frac{1}{2} - \frac{M_{\mathcal{L}}}{\sqrt{\mu_{\mathcal{L}}^{\text{A}}}} \quad \text{(see Equation (75) for the definition of } M_{\mathcal{L}}\text{)}. \tag{135}$$

Let the tuple $(g_t, \nu, \mu_{\mathcal{L}})$ refers to either $(g_t^{\text{FW}}, \nu^{\text{FW}}, \mu_{\mathcal{L}}^{int})$ for case (I) where the algorithm is SP-FW , or $(g_t^{\text{PFW}}, \nu^{\text{PFW}}, \mu_{\mathcal{L}}^{\text{A}})$ for case (P) where the algorithm is SP-AFW,

If $\nu > 0$, then the suboptimality h_t of the iterates of the algorithm with step size $\gamma_t = \min(\gamma_{\max}, \frac{\nu}{2C_{\mathcal{L}}} g_t)$ decreases geometrically[4] as,

$$h_t \leq P_{\mathcal{L}} \sqrt{2w_0} (1 - \rho_{\mathcal{L}})^{k(t)/2} \tag{136}$$

where $\rho_{\mathcal{L}} := \frac{\nu^2 \mu_{\mathcal{L}}}{2C_{\mathcal{L}}}$ and $k(t)$ is the number of non-drop step after t steps. For SP-FW, $k(t) = t$ and for SP-AFW, $k(t) \geq t/3$. Moreover we can also upper bound the minimum gap observed, for all $T \in \mathbb{N}$

$$\min_{t \leq T} g_t \leq \frac{2\max\{\sqrt{C_{\mathcal{L}}}, \sqrt{w_0}(1 - \rho_{\mathcal{L}})^{k(T)/2}\}}{\nu} \sqrt{w_0} \, (1 - \rho_{\mathcal{L}})^{k(T)/2}. \tag{137}$$

The Theorem 1 statement can be deduced from this theorem using the lower and upper bounds on the affine invariant constant of this statement. More precisely, one can upper bound $C_{\mathcal{L}}$, $M_{\mathcal{L}}$, $P_{\mathcal{L}}$ respectively with Propositions 9, 13 and 14 and lower bound $\mu_{\mathcal{L}}^{int}$ and $\mu_{\mathcal{L}}^{\text{A}}$ respectively with Proposition 12 and Equation (72).[5] If we apply these bounds to the rate multipliers in (135), it gives the bigger rate multipliers ν stated in Theorem 1.

Proof. We uses the Lemma 23 giving a geometric scheme, with a straightforward recurrence we prove that,

$$w_t \leq w_0 (1 - \rho_{\mathcal{L}})^{k(t)}, \tag{138}$$

[4] For a non-drop step one can use Corollary 22 to get the better rate on h_t losing the square root but with a potentially worse constant $h_t \leq \frac{2\max\{C_{\mathcal{L}}, w_0(1-\rho_{\mathcal{L}})^{k(t)}\}}{\mu_{\mathcal{L}}^{\text{A}}(\nu^{\text{PFW}})^2} w_0 (1 - \rho_{\mathcal{L}})^{k(t)}$.

[5] Note that only the definition of δ_μ in Theorem 1 for case (P) requires to use the Euclidean norm (because inequality (72) with the pyramidal width only holds for the Euclidean norm). On the other hand, any norm could be used for (separately) bounding $C_{\mathcal{L}}$, $M_{\mathcal{L}}$ and $P_{\mathcal{L}}$.

where $k(t)$ is the number of non-drop step steps. This number is equal to t for the SP-FW algorithm and it is lower bounded by $t/3$ for the SP-AFW algorithm (see Section A Equation (38)). Then by using Proposition (15) relating h_t and the square root of w_t we get the first statement of the theorem,

$$h_t \leq P_{\mathcal{L}} \sqrt{2w_0}(1-\rho_{\mathcal{L}})^{k(t)/2}. \tag{139}$$

To prove the second statement of the theorem we just use Theorem 21 for the last *non-drop step* after T iterations (let us assume it was at step t_0),

$$g_{t_0} \leq \frac{2\max\{\sqrt{C_{\mathcal{L}}}, \sqrt{w_0}(1-\rho_{\mathcal{L}})^{k(t_0)/2}\}}{\nu} \sqrt{w_0}(1-\rho_{\mathcal{L}})^{k(t_0)/2} \tag{140}$$

$$= \frac{2\max\{\sqrt{C_{\mathcal{L}}}, \sqrt{w_0}(1-\rho_{\mathcal{L}})^{k(T)/2}\}}{\nu} \sqrt{w_0}(1-\rho_{\mathcal{L}})^{k(T)/2}. \qquad (\text{because } k(t_0)=k(T)) \tag{141}$$

The minimum of the gaps observed is smaller than the gap at time t_0 then,

$$\min_{t\leq T} g_t \leq g_{t_0} \leq \frac{2\max\{\sqrt{C_{\mathcal{L}}}, \sqrt{w_0}(1-\rho_{\mathcal{L}})^{k(T)/2}\}}{\nu} \sqrt{w_0}(1-\rho_{\mathcal{L}})^{k(T)/2}. \tag{142}$$

□

The affine invariant formulation with the universal step size $\gamma_t = \min\left\{\gamma_{\max}, \frac{2}{2+k(t)}\right\}$ of Theorem 1 also follows from Lemma 23 by re-using standard FW proof patterns.

Theorem 25. *Let $\mathcal{L}$ be a strongly convex-concave function with a finite smoothness constant $C_{\mathcal{L}}$, a positive interior strong convex-concavity constant $\mu_{\mathcal{L}}^{int}$ (66) or a positive geometric strong convex-concavity $\mu_{\mathcal{L}}^{\mathrm{A}}$ (71). Let us also define the rate multipliers ν as*

$$\nu^{\mathrm{FW}} := 1 - \frac{M_{\mathcal{L}}}{\sqrt{\mu_{\mathcal{L}}^{int}}} \quad \text{and} \quad \tilde{\nu}^{\mathrm{PFW}} := 1 - \frac{M_{\mathcal{L}}}{\sqrt{\mu_{\mathcal{L}}^{\mathrm{A}}}} \quad (\text{see Equation (75) for the definition of } M_{\mathcal{L}}). \tag{143}$$

Let ν refers to either ν^{FW} for case (I) where the algorithm is SP-FW, or $\tilde{\nu}^{\mathrm{PFW}}$ for case (P) where the algorithm is SP-AFW,

If $\nu > \frac{1}{2}$, then the suboptimality w_t of the iterates of the algorithm with universal step size $\gamma_t = \min\left\{\gamma_{\max}, \frac{2}{2+k(t)}\right\}$ (see Equation (34) for more details about $\gamma_{\max}$) has the following decreasing upper bound:

$$w_t \leq \frac{C}{2+k(t)} \tag{144}$$

where $C = 2\max\left(w_0, \frac{2C_{\mathcal{L}}}{2\nu-1}\right)$ and $k(t)$ is the number of non-drop step after t steps. For SP-FW, $k(t)=t$ and for SP-AFW, $k(t) \geq t/3$. Moreover we can also upper bound the minimum FW gap observed for $T \geq 1$,

$$\min_{t\leq T} g_t^{\mathrm{FW}} \leq \frac{5C}{\nu(k(T)+1)}. \tag{145}$$

Note that in this theorem the constant $\tilde{\nu}^{\mathrm{PFW}}$ is slightly different from the constant ν^{PFW} in Theorem 24.

Proof. We can put both the recurrence (115) for the SP-FW algorithm and the recurrence (117) for the SP-AFW algorithm (from the proof of Lemma 20) in the following form by using our unified notation introduced in the theorem statement:

$$w_{t+1} \leq w_t - \gamma_t \nu g_t + \gamma_t^2 C_{\mathcal{L}}. \tag{146}$$

Note that the gap g_t is the one defined in Equation (88) and depends on the algorithm. Let (ν) to stand respectively for (ν^{FW}) for SP-FW (case (I)) or $(\tilde{\nu}^{\mathrm{PFW}})$ for SP-AFW (case (P)). With this notation, the inequality $g_t \geq w_t$ leads to,

$$w_{t+1} \leq w_t\,(1-\nu\gamma_t) + \gamma_t^2 C_{\mathcal{L}}. \tag{147}$$

Our goal is to show by induction that

$$w_t \leq \frac{C}{2+k(t)} \qquad \text{where } C := 2\max\left(w_0, \frac{2C_{\mathcal{L}}}{2\nu-1}\right). \tag{$\star$}$$

Let us first define the convex function $f_t : \gamma \mapsto w_t\,(1-\nu\gamma) + \gamma^2 C_{\mathcal{L}}$. We will show that under ($\star$), the function f_t has the following property:

$$f_t\left(\frac{2}{2+k(t)}\right) \leq \frac{C}{3+k(t)}. \tag{148}$$

This property is due to a simple inequality on integers; let $k = k(t)$, from the crucial induction assumption, we get:

$$f_t\left(\frac{2}{2+k}\right) = w_t\frac{2+k-2\nu}{2+k} + \frac{4}{(2+k)^2}C_{\mathcal{L}} \leq \frac{C}{3+k}\left[\frac{(3+k)(k+1-(2\nu-1)+\frac{4C_{\mathcal{L}}}{C})}{(2+k)^2}\right], \tag{149}$$

but $(2\nu-1) \geq \frac{4C_{\mathcal{L}}}{C}$ and $(3+k)(1+k) < (2+k)(2+k)$ for any k, thus

$$f_t\left(\frac{2}{2+k}\right) \leq \frac{C}{3+k}. \tag{150}$$

Equation (150) is crucial for the inductive step of our recurrence.

- Hypothesis ($\star$) is true for $t=0$ because $k(0)=0$.
- Now let us assume that ($\star$) is true for a $t \in \mathbb{N}$. We set the stepsize $\gamma_t := \min\left\{\gamma_{\max}, \frac{2}{2+k(t)}\right\}$.

  If $k(t+1) = k(t)+1$, it means that $\gamma_t = \frac{2}{2+k(t)}$ and then by (147) and (150),

  $$w_{t+1} \leq f_t\left(\frac{2}{2+k(t)}\right) \leq \frac{C}{3+k(t)} = \frac{C}{2+k(t+1)}. \tag{151}$$

  If $k(t+1) = k(t)$, then it means that $0 \leq \gamma_t < \frac{2}{2+k(t)}$. Hence, the convexity of the function f_t leads us to the inequality

  $$\begin{aligned} w_{t+1} \leq f_t(\gamma_t) &\leq \max\left\{f_t(0), f_t\left(\frac{2}{2+k(t)}\right)\right\} \\ &= \max\left\{w_t, f_t\left(\frac{2}{2+k(t)}\right)\right\} &(152)\\ &\leq \max\left\{\frac{C}{2+k(t)}, \frac{C}{3+k(t)}\right\} &(153)\\ &\leq \frac{C}{2+k(t)}. &(154)\end{aligned}$$

  where we used (150) and the induction hypothesis ($\star$) to get the penultimate inequality (152). Since we assumed that $k(t+1) = k(t)$, we get

  $$w_{t+1} \leq \frac{C}{2+k(t+1)}, \tag{155}$$

  completing the induction proof for (144).

In case (I), $k(t) = t$ and in case (P), $k(t) \geq t/3$ (see Equation (38)), leading us to the first statement of our theorem.

The proof of the second statement is inspired by the proof of Theorem C.3 from (Lacoste-Julien et al., 2013).

With the same notation as the proof of Lemma 20, we start from Equation (146) where we isolated the gap g_t to get the crucial inequality

$$g_t \leq \frac{w_t - w_{t+1}}{\nu \gamma_t} + \gamma_t \frac{C_{\mathcal{L}}}{\nu}. \tag{156}$$

Since the gap g_t is the one depending on the algorithm defined by $g_t := \left\langle -\boldsymbol{r}^{(t)}, \boldsymbol{d}^{(t)} \right\rangle$, we have $g_t = g_t^{\mathrm{FW}}$ for SP-FW and $g_t = \max\left(g_t^{\mathrm{FW}}, g_t^{\mathrm{A}}\right) \geq g_t^{\mathrm{FW}}$ for SP-AFW. Thus,

$$g_t^{\mathrm{FW}} \leq g_t \leq \frac{w_t - w_{t+1}}{\nu \gamma_t} + \gamma_t \frac{C_{\mathcal{L}}}{\nu}. \tag{157}$$

In the following in order not to be too heavy with notation we will work with de FW gap and note g_t for g_t^{FW}.

The proof idea is to take a convex combination of the inequality (157) to obtain a new upper-bound on a convex combination of the gaps computed from step 0 to step T. Let us introduce the convex combination weight $\rho_t := \frac{\gamma_t \cdot k(t)(k(t)+2)}{S_T}$ where $k(t)$ is the number of non-drop steps after t steps and S_T is the normalization factor. Let us also call $N_T := \{t \leq T \,|\, t \text{ is a non-drop step}\}$. Taking the convex combination of (157), we get

$$\sum_{t=0}^{T} \rho_t g_t \leq \sum_{t=0}^{T} \rho_t \frac{w_t - w_{t+1}}{\nu \gamma_t} + \sum_{t=0}^{T} \rho_t \gamma_t \frac{C_{\mathcal{L}}}{\nu}. \tag{158}$$

By regrouping the terms and ignoring the negative term, we get

$$\sum_{t=0}^{T} \rho_t g_t \leq \frac{w_0 \rho_0}{\nu \gamma_0} + \frac{1}{\nu} \sum_{t=0}^{T-1} w_{t+1} \left(\frac{\rho_{t+1}}{\gamma_{t+1}} - \frac{\rho_t}{\gamma_t} \right) + \sum_{t=0}^{T} \rho_t \gamma_t \frac{C_{\mathcal{L}}}{\nu}. \tag{159}$$

By definition $\frac{\rho_t}{\gamma_t} := \frac{k(t)(k(t)+2)}{S_T}$ and notice that $\rho_0 = 0$. We now consider two possibilities: if γ_t is a drop step, then $k(t+1) = k(t)$ and so

$$\frac{\rho_{t+1}}{\gamma_{t+1}} - \frac{\rho_t}{\gamma_t} = 0. \tag{160}$$

If γ_t is a non-drop step, then $k(t+1) = k(t) + 1$ and thus we have

$$\frac{\rho_{t+1}}{\gamma_{t+1}} - \frac{\rho_t}{\gamma_t} = \frac{(k(t)+1)(k(t)+3)}{S_T} - \frac{k(t)(k(t)+2)}{S_T} \tag{161}$$

$$= \frac{2k(t)+3}{S_T}. \tag{162}$$

As $\gamma_t \leq \frac{2}{k(t)+2}$, we also have $\rho_t \gamma_t \leq \frac{4k(t)}{S_T(k(t)+2))}$. The normalization factor S_T to define a convex combination is equal to

$$S_T := \sum_{u=0}^{T} \gamma_u \cdot k(u)(k(u)+2) \geq \sum_{\substack{u=0 \\ u \in N_T}}^{T} \frac{2}{2+k(u)} \cdot k(u)(k(u)+2) = \sum_{k=0}^{k(T)} 2k = k(T)(K(T)+1). \tag{163}$$

Plugging this and (162) in the inequality (159) with the rate ($\star$) shown by induction gives us,

$$\sum_{t=0}^{T} \rho_t g_t \leq 0 + \frac{1}{\nu} \sum_{\substack{t=0 \\ t \in N_T}}^{T-1} \frac{C}{2 + k(t+1)} \frac{2k(t)+3}{k(T)(k(T)+1)} + \sum_{t=0}^{T} \frac{4k(t)}{k(T)(k(T)+1)(k(t)+2)} \frac{C_{\mathcal{L}}}{\nu} \tag{164}$$

$$\leq \frac{2C}{\nu} \frac{1}{k(T)(k(T)+1)} \Big(\sum_{\substack{t=0 \\ t \in N_T}}^{T-1} 1 + \frac{2C_{\mathcal{L}}}{C} \sum_{t=1}^{T} 1 \Big) \tag{165}$$

$$\leq \frac{2C}{\nu(k(T)+1)} (1 + \frac{\nu}{2} \frac{T}{k(T)}) \leq \frac{5C}{\nu(k(T)+1)}, \tag{166}$$

by using $k(T) \geq T/3$. Finally, the minimum of the gaps is always smaller than any convex combination, so we can conclude that (for $T \geq 1$):

$$\min_{0 \leq t \leq T} g_t \leq \frac{5C}{\nu(k(T)+1)}. \tag{167}$$

□

E Strongly convex sets

In this section, we are going to prove that the function $\boldsymbol{s}(\cdot)$ is Lipschitz continuous when the sets $\mathcal{X}$ and $\mathcal{Y}$ are strongly convex and when the norm of the two gradient components are uniformly lower bounded. We will also give the details of the convergence rate proof for the strongly convex sets situation. Our proof uses similar arguments as Dunn (1979, Theorem 3.4 and 3.6).

Theorem' 3. *Let $\mathcal{X}$ and $\mathcal{Y}$ be β-strongly convex sets. If $\min(\|\nabla_x L(\boldsymbol{z})\|_{\mathcal{X}^*}, \|\nabla_y L(\boldsymbol{z})\|_{\mathcal{Y}^*}) \geq \delta > 0$ for all $\boldsymbol{z} \in \mathcal{X} \times \mathcal{Y}$, then the oracle function $\boldsymbol{z} \mapsto \boldsymbol{s}(\boldsymbol{z}) := \arg\min_{\boldsymbol{s} \in \mathcal{X} \times \mathcal{Y}} \langle \boldsymbol{s}, F(\boldsymbol{z}) \rangle$ is well defined and is $\frac{4L}{\delta\beta}$-Lipschitz continuous (using the norm $\|(\boldsymbol{x}, \boldsymbol{y})\|_{\mathcal{X} \times \mathcal{Y}} := \|\boldsymbol{x}\|_{\mathcal{X}} + \|\boldsymbol{y}\|_{\mathcal{Y}}$), where $F(\boldsymbol{z}) := (\nabla_x \mathcal{L}(\boldsymbol{z}), -\nabla_y \mathcal{L}(\boldsymbol{z}))$.*

Proof. First note that since the sets are strongly convex, the minimum is reached at a unique point. Then, we introduce the following lemma which can be used to show that each component of the gradient is Lipschitz continuous irrespective of the other set.

Lemma 26. *Let $F_x : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ be a L-Lipschitz continuous function (i.e. $\|F_x(\boldsymbol{z}) - F_x(\boldsymbol{z}')\|_{\mathcal{X}^*} \leq L\|\boldsymbol{z} - \boldsymbol{z}'\|_{\mathcal{X} \times \mathcal{A}}$) and $\mathcal{X}$ a β-strongly convex set. If $\forall \boldsymbol{z} \in \mathcal{X} \times \mathcal{A}$, $\|F_x(\boldsymbol{z})\|_{\mathcal{X}^*} \geq \delta > 0$, then $\boldsymbol{s}_x : \boldsymbol{z} \mapsto \arg\min_{\boldsymbol{s} \in \mathcal{X}} \langle \boldsymbol{s}, F_x(\boldsymbol{z}) \rangle$ is $\frac{2L}{\delta\beta}$-Lipschitz continuous.*

Proof. Let $\boldsymbol{z}$, $\boldsymbol{z}' \in \mathcal{X} \times \mathcal{A}$ and let $\bar{\boldsymbol{x}} = \frac{\boldsymbol{s}_x(\boldsymbol{z}) + \boldsymbol{s}_x(\boldsymbol{z}')}{2}$, then

$$\begin{aligned} \langle \boldsymbol{s}_x(\boldsymbol{z}) - \boldsymbol{s}_x(\boldsymbol{z}'), -F_x(\boldsymbol{z}) \rangle &= 2 \langle \boldsymbol{s}_x(\boldsymbol{z}) - \bar{\boldsymbol{x}}, -F_x(\boldsymbol{z}) \rangle \\ &\geq 2 \langle \boldsymbol{x} - \bar{\boldsymbol{x}}, -F_x(\boldsymbol{z}) \rangle . \quad \forall \boldsymbol{x} \in \mathcal{X} \quad \text{(by definition of } s_x) \end{aligned} \tag{168}$$

Now (168) holds for any $\boldsymbol{x} \in B_\beta \left(\frac{1}{2}, \boldsymbol{s}_x(\boldsymbol{z}), \boldsymbol{s}_x(\boldsymbol{z}') \right)$ as this set is included in $\mathcal{X}$ by β-strong convexity of $\mathcal{X}$. Then since $\bar{\boldsymbol{x}}$ is the center of $B_\beta \left(\frac{1}{2}, \boldsymbol{s}_x(\boldsymbol{z}), \boldsymbol{s}_x(\boldsymbol{z}') \right)$, we can choose a $\boldsymbol{x}$ in this ball such that $\boldsymbol{x} - \bar{\boldsymbol{x}}$ is in the direction which achieves the dual norm of $-F_x(\boldsymbol{z})$.[6] More specifically, we have that:

$$\| - F_x(\boldsymbol{z}) \|_{\mathcal{X}^*} = \sup_{\|\boldsymbol{v}\|_{\mathcal{X}} \leq 1} \langle -F_x(\boldsymbol{z}), \boldsymbol{v} \rangle .$$

[6] For the Euclidean norm, we choose $\boldsymbol{x} - \bar{\boldsymbol{x}}$ proportional to $-F_x$; but for general norms, it could be a different direction.

As we are in finite dimensions, this supremum is achieved by some vector $\boldsymbol{v}$. So choose $\boldsymbol{x} := \bar{\boldsymbol{x}} + \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_{\mathcal{X}}} \frac{\beta}{8} \|\boldsymbol{s}_x(\boldsymbol{z}) - \boldsymbol{s}_x(\boldsymbol{z}')\|_{\mathcal{X}}^2 \in B_\beta\left(\frac{1}{2}, \boldsymbol{s}_x(\boldsymbol{z}), \boldsymbol{s}_x(\boldsymbol{z}')\right)$ and plug it in (168):

$$\left\langle \boldsymbol{s}_x(\boldsymbol{z}) - \boldsymbol{s}_x(\boldsymbol{z}'), -F_x(\boldsymbol{z}) \right\rangle \geq \frac{\beta \|\boldsymbol{s}_x(\boldsymbol{z}) - \boldsymbol{s}_x(\boldsymbol{z}')\|_{\mathcal{X}}^2}{4\|\boldsymbol{v}\|_{\mathcal{X}}} \langle \boldsymbol{v}, -F_x(\boldsymbol{z}) \rangle \tag{169}$$

$$= \frac{\beta \|\boldsymbol{s}_x(\boldsymbol{z}) - \boldsymbol{s}_x(\boldsymbol{z}')\|_{\mathcal{X}}^2}{4\|\boldsymbol{v}\|_{\mathcal{X}}} \|F_x(\boldsymbol{z})\|_{\mathcal{X}^*} \tag{170}$$

$$\geq \frac{\beta}{4} \|\boldsymbol{s}_x(\boldsymbol{z}) - \boldsymbol{s}_x(\boldsymbol{z}')\|_{\mathcal{X}}^2 \|F_x(\boldsymbol{z})\|_{\mathcal{X}^*}. \tag{171}$$

Switching $\boldsymbol{z}$ and $\boldsymbol{z}'$ and using a similar argument, we get,

$$\left\langle \boldsymbol{s}_x(\boldsymbol{z}') - \boldsymbol{s}_x(\boldsymbol{z}), -F_x(\boldsymbol{z}') \right\rangle \geq \frac{\beta}{4} \|\boldsymbol{s}_x(\boldsymbol{z}) - \boldsymbol{s}_x(\boldsymbol{z}')\|_{\mathcal{X}}^2 \|F_x(\boldsymbol{z}')\|_{\mathcal{X}^*}. \tag{172}$$

Hence summing (171) and (172),

$$\begin{aligned}
\frac{\beta}{4} \left(\|F_x(\boldsymbol{z})\|_{X^*} + \|F_x(\boldsymbol{z}')\|_{\mathcal{X}^*}\right) \|\boldsymbol{s}_x(\boldsymbol{z}) - \boldsymbol{s}_x(\boldsymbol{z}')\|_{\mathcal{X}}^2 &\leq \left\langle \boldsymbol{s}_x(\boldsymbol{z}) - \boldsymbol{s}_x(\boldsymbol{z}'), F_x(\boldsymbol{z}') - F_x(\boldsymbol{z}) \right\rangle \\
&\leq \|\boldsymbol{s}_x(\boldsymbol{z}) - \boldsymbol{s}_x(\boldsymbol{z}')\|_{\mathcal{X}} \|F_x(\boldsymbol{z}') - F_x(\boldsymbol{z})\|_{\mathcal{X}^*} \qquad (173) \\
&\leq L \|\boldsymbol{s}_x(\boldsymbol{z}) - \boldsymbol{s}_x(\boldsymbol{z}')\|_{\mathcal{X}} \|\boldsymbol{z}' - \boldsymbol{z}\|_{\mathcal{X}\times\mathcal{Y}} \quad \text{(Lip. cty. of } F_x\text{)}
\end{aligned}$$

and finally

$$\|\boldsymbol{s}_x(\boldsymbol{z}) - \boldsymbol{s}_x(\boldsymbol{z}')\|_{\mathcal{X}} \leq \frac{4L}{\beta \left(\|F_x(\boldsymbol{z})\|_{\mathcal{X}^*} + \|F_x(\boldsymbol{z}')\|_{\mathcal{X}^*}\right)} \|\boldsymbol{z} - \boldsymbol{z}'\|_{\mathcal{X}\times\mathcal{Y}} \leq \frac{2L}{\delta\beta} \|\boldsymbol{z} - \boldsymbol{z}'\|_{\mathcal{X}\times\mathcal{Y}}. \tag{174}$$

□

To prove our theorem, we will notice that for the saddle point setup, the oracle function $\boldsymbol{s}(\cdot) := \arg\min_{\boldsymbol{s}\in\mathcal{X}\times\mathcal{Y}} \langle \boldsymbol{s}, F(\cdot) \rangle$ can be decomposed as $\boldsymbol{s}(\cdot) = (\boldsymbol{s}_x(\cdot), \boldsymbol{s}_y(\cdot))$ where $\boldsymbol{s}_x(\cdot) := \arg\min_{\boldsymbol{s}\in\mathcal{X}} \langle \boldsymbol{s}, F_x(\cdot) \rangle$ and $\boldsymbol{s}_y(\cdot) := \arg\min_{\boldsymbol{s}\in\mathcal{Y}} \langle \boldsymbol{s}, F_y(\cdot) \rangle$. Then applying our lemma, the function $\boldsymbol{s}_x(\cdot)$ is Lipschitz continuous. The same way $\boldsymbol{s}_y(\cdot)$ is Lipschitz continuous. Then, for all $\boldsymbol{z}, \boldsymbol{z}'$ in $\mathcal{X} \times \mathcal{Y}$

$$\|\boldsymbol{s}(\boldsymbol{z}) - \boldsymbol{s}(\boldsymbol{z}')\|_{\mathcal{X}\times\mathcal{Y}} = \|\boldsymbol{s}_x(\boldsymbol{z}) - \boldsymbol{s}_x(\boldsymbol{z}')\|_{\mathcal{X}} + \|\boldsymbol{s}_y(\boldsymbol{z}) - \boldsymbol{s}_y(\boldsymbol{z}')\|_{\mathcal{Y}} \leq \frac{4L}{\delta\beta} \|\boldsymbol{z} - \boldsymbol{z}'\|_{\mathcal{X}\times\mathcal{Y}}, \tag{175}$$

which gives the definition of the Lipschitz continuity of our function and proves the theorem. □

In this theorem, we introduced the function F. This function is monotone in the following sense:

$$\forall \boldsymbol{z}, \boldsymbol{z}' \quad \left\langle \boldsymbol{z} - \boldsymbol{z}', F(\boldsymbol{z}) - F(\boldsymbol{z}') \right\rangle \geq 0. \tag{176}$$

Actually this property follows directly from the convexity of $\mathcal{L}(\cdot, \boldsymbol{y})$ and the concavity of $\mathcal{L}(\boldsymbol{x}, \cdot)$. We can also prove that when the sets $\mathcal{X}$ and $\mathcal{Y}$ are strongly convex and when the gradient is uniformly lower bounded, we can relate the gap and the distance between $\boldsymbol{z}^{(t)}$ and $\boldsymbol{s}^{(t)}$.

Lemma 27. *If $\mathcal{X}$ is a β-strongly convex set and if $\|\nabla f\|_{\mathcal{X}^*}$ is uniformly lower bounded by δ on $\mathcal{X}$, then*

$$\max_{\boldsymbol{s}\in\mathcal{X}} \langle \boldsymbol{s} - \boldsymbol{x}, -\nabla f(\boldsymbol{x}) \rangle \geq \frac{\beta}{4} \delta \|\boldsymbol{s}(\boldsymbol{x}) - \boldsymbol{x}\|^2, \tag{177}$$

where $\boldsymbol{s}(\boldsymbol{x}) = \arg\max_{\boldsymbol{s}\in\mathcal{X}} \langle \boldsymbol{s} - \boldsymbol{x}, -\nabla f(\boldsymbol{x}) \rangle$.

Proof. Let $\boldsymbol{x}$ and $\boldsymbol{s}(\boldsymbol{x})$ be in $\mathcal{X}$. We have $B_\beta\left(\frac{1}{2}, \boldsymbol{s}(\boldsymbol{x}), \boldsymbol{x}\right) \subset \mathcal{X}$ by β-strong convexity. So as in the proof of Lemma 26, let $\boldsymbol{v}$ be the vector such that $\|\boldsymbol{v}\|_{\mathcal{X}} \leq 1$ and $\langle -\nabla f(\boldsymbol{x}), \boldsymbol{v}\rangle = \|\nabla f(\boldsymbol{x})\|_{\mathcal{X}^*}$. Let

$$\bar{\boldsymbol{s}} := \frac{\boldsymbol{s}(\boldsymbol{x}) + \boldsymbol{x}}{2} + \frac{\beta}{8}\|\boldsymbol{s}(\boldsymbol{x}) - \boldsymbol{x}\|^2 \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_{\mathcal{X}}} \in \mathcal{X}. \tag{178}$$

Then

$$\begin{aligned}
\langle \boldsymbol{s}(\boldsymbol{x}) - \boldsymbol{x}, -\nabla f(\boldsymbol{x})\rangle &\geq \langle \bar{\boldsymbol{s}} - \boldsymbol{x}, -\nabla f(\boldsymbol{x})\rangle \\
&= \frac{1}{2}\langle \boldsymbol{s}(\boldsymbol{x}) - \boldsymbol{x}, -\nabla f(\boldsymbol{x})\rangle + \frac{\beta}{8}\|\boldsymbol{s}(\boldsymbol{x}) - \boldsymbol{x}\|^2 \frac{\|\nabla f(\boldsymbol{x})\|_{\mathcal{X}^*}}{\|\boldsymbol{v}\|_{\mathcal{X}}} \\
&\geq \frac{1}{2}\langle \boldsymbol{s}(\boldsymbol{x}) - \boldsymbol{x}, -\nabla f(\boldsymbol{x})\rangle + \frac{\beta}{8}\delta\|\boldsymbol{s}(\boldsymbol{x}) - \boldsymbol{x}\|^2
\end{aligned} \tag{179}$$

which leads us to the desired result. □

From this lemma, under the assumption that $\min(\|\nabla_x L(\boldsymbol{z})\|_{\mathcal{X}^*}, \|\nabla_y L(\boldsymbol{z})\|_{\mathcal{Y}^*}) \geq \delta\ \forall \boldsymbol{z} \in \mathcal{X}\times\mathcal{Y}$, it directly follows that

$$\begin{aligned}
g_t^{\mathrm{FW}} = g_t^{(x)} + g_t^{(y)} &\geq \frac{\beta}{4}\delta\left(\|\boldsymbol{s}_x^{(t)} - \boldsymbol{x}^{(t)}\|_{\mathcal{X}}^2 + \|\boldsymbol{s}_y^{(t)} - \boldsymbol{y}^{(t)}\|_{\mathcal{Y}}^2\right) \\
&\geq \frac{\beta}{8}\delta\left(\|\boldsymbol{s}_x^{(t)} - \boldsymbol{x}^{(t)}\|_{\mathcal{X}} + \|\boldsymbol{s}_y^{(t)} - \boldsymbol{y}^{(t)}\|_{\mathcal{Y}}\right)^2 = \frac{\beta}{8}\delta\|\boldsymbol{s}^{(t)} - \boldsymbol{z}^{(t)}\|_{\mathcal{X}\times\mathcal{Y}}^2.
\end{aligned} \tag{180}$$

Now we recall the convergence theorem for strongly convex sets from the main text, Theorem 4:

Theorem' 4. *Let $\mathcal{L}$ be a convex-concave function and $\mathcal{X}$ and $\mathcal{Y}$ two compact β-strongly convex sets. Assume that the gradient of $\mathcal{L}$ is L-Lipschitz continuous and that there exists $\delta > 0$ such that $\min(\|\nabla_x L(\boldsymbol{z})\|_{\mathcal{X}^*}, \|\nabla_y L(\boldsymbol{z})\|_{\mathcal{Y}^*}) \geq \delta\ \ \forall \boldsymbol{z} \in \mathcal{X}\times\mathcal{Y}$. Set $C_\delta := 2L + \frac{8L^2}{\beta\delta}$. Then the gap g_t^{FW} (7) of the SP-FW algorithm with step size $\gamma_t = \frac{g_t^{\mathrm{FW}}}{\|\boldsymbol{s}^{(t)} - \boldsymbol{z}^{(t)}\|^2 C_\delta}$ converges linearly as*

$$g_t^{\mathrm{FW}} \leq g_0\,(1-\rho)^t \tag{181}$$

where $\rho := \frac{\beta\delta}{16 C_\delta}$. The initial gap g_0 is cheaply computed during the first step of the SP-FW algorithm. Alternatively, one can use the following upper bound to get uniform guarantees:

$$g_0 \leq \sup_{\boldsymbol{z}\in\mathcal{X}\times\mathcal{Y}} \|\nabla_x \mathcal{L}(\boldsymbol{z})\|_{\mathcal{X}^*} D_{\mathcal{X}} + \sup_{\boldsymbol{z}\in\mathcal{X}\times\mathcal{Y}} \|\nabla_y \mathcal{L}(\boldsymbol{z})\|_{\mathcal{Y}^*} D_{\mathcal{Y}}. \tag{182}$$

Proof. We compute the following relation on the gap:

$$\begin{aligned}
g_{t+1} &= \left\langle \boldsymbol{z}^{(t+1)} - \boldsymbol{s}^{(t+1)}, F(\boldsymbol{z}^{(t+1)})\right\rangle \\
&= \left\langle \boldsymbol{z}^{(t)} - \boldsymbol{s}^{(t+1)}, F(\boldsymbol{z}^{(t+1)})\right\rangle + \gamma_t\left\langle \boldsymbol{s}^{(t)} - \boldsymbol{z}^{(t)}, F(\boldsymbol{z}^{(t+1)})\right\rangle \\
&= \left\langle \boldsymbol{z}^{(t)} - \boldsymbol{s}^{(t+1)}, F(\boldsymbol{z}^{(t)})\right\rangle + \left\langle \boldsymbol{z}^{(t)} - \boldsymbol{s}^{(t+1)}, F(\boldsymbol{z}^{(t+1)}) - F(\boldsymbol{z}^{(t)})\right\rangle \\
&\quad + \gamma_t\left\langle \boldsymbol{s}^{(t)} - \boldsymbol{z}^{(t)}, F(\boldsymbol{z}^{(t)})\right\rangle + \gamma_t\left\langle \boldsymbol{s}^{(t)} - \boldsymbol{z}^{(t)}, F(\boldsymbol{z}^{(t+1)}) - F(\boldsymbol{z}^{(t)})\right\rangle \\
&\leq \left\langle \boldsymbol{z}^{(t)} - \boldsymbol{s}^{(t+1)}, F(\boldsymbol{z}^{(t)})\right\rangle + \left\langle \boldsymbol{z}^{(t)} - \boldsymbol{s}^{(t+1)}, F(\boldsymbol{z}^{(t+1)}) - F(\boldsymbol{z}^{(t)})\right\rangle \\
&\quad + \gamma_t\left\langle \boldsymbol{s}^{(t)} - \boldsymbol{z}^{(t)}, F(\boldsymbol{z}^{(t)})\right\rangle + \gamma_t^2\|\boldsymbol{s}^{(t)} - \boldsymbol{z}^{(t)}\|^2 L
\end{aligned} \tag{183}$$

where in the last line we used the fact that the function $F(\cdot)$ is Lipschitz continuous. Then using that $\left\langle \boldsymbol{z}^{(t)} - \boldsymbol{s}^{(t+1)}, F(\boldsymbol{z}^{(t)}) \right\rangle \leq \left\langle \boldsymbol{z}^{(t)} - \boldsymbol{s}^{(t)}, F(\boldsymbol{z}^{(t)}) \right\rangle$ (by definition of $\boldsymbol{s}^{(t)}$), we get

$$\begin{aligned} g_{t+1} &\leq g_t(1-\gamma_t) + \left\langle \boldsymbol{z}^{(t)} - \boldsymbol{s}^{(t+1)}, F(\boldsymbol{z}^{(t+1)}) - F(\boldsymbol{z}^{(t)}) \right\rangle + \gamma_t^2 \|\boldsymbol{s}^{(t)} - \boldsymbol{z}^{(t)}\|^2 L \\ &\leq g_t(1-\gamma_t) + \left\langle \boldsymbol{s}^{(t)} - \boldsymbol{s}^{(t+1)}, F(\boldsymbol{z}^{(t+1)}) - F(\boldsymbol{z}^{(t)}) \right\rangle + \gamma_t^2 \|\boldsymbol{s}^{(t)} - \boldsymbol{z}^{(t)}\|^2 L. \end{aligned} \tag{184}$$

The last line uses the fact that F is monotone by convexity (Equation (176)). Finally, using once again the Lipschitz continuity of F and the one of $\boldsymbol{s}(\cdot)$ (by Theorem 3), we get

$$\begin{aligned} \left\langle \boldsymbol{s}^{(t)} - \boldsymbol{s}^{(t+1)}, F(\boldsymbol{z}^{(t+1)}) - F(\boldsymbol{z}^{(t)}) \right\rangle &\leq \|\boldsymbol{s}^{(t)} - \boldsymbol{s}^{(t+1)}\| L \|\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)}\| \\ &\leq \frac{4L^2}{\beta\delta} \|\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)}\|^2 \qquad \text{(Lipschitz continuity of } \boldsymbol{s}) \\ &= \frac{4L^2}{\beta\delta} \gamma_t^2 \|\boldsymbol{s}^{(t)} - \boldsymbol{z}^{(t)}\|^2. \end{aligned} \tag{185}$$

Combining (185) with (184), we get

$$g_{t+1} \leq g_t(1-\gamma_t) + \gamma_t^2 \|\boldsymbol{s}^{(t)} - \boldsymbol{z}^{(t)}\|^2 \frac{C_\delta}{2} \qquad \text{where} \qquad C_\delta := 2L + \frac{8L^2}{\beta\delta}. \tag{186}$$

Thus by setting the step size $\gamma_t = \frac{g_t}{\|\boldsymbol{s}^{(t)} - \boldsymbol{z}^{(t)}\|^2 C_\delta}$, we get

$$g_{t+1} \leq g_t - \frac{g_t}{2C_\delta} \left(\frac{g_t}{\|\boldsymbol{s}^{(t)} - \boldsymbol{z}^{(t)}\|^2} \right) \leq g_t \left(1 - \frac{\beta\delta}{16 C_\delta} \right), \tag{187}$$

using the fact that the gap is lower bounded by a constant times the square of the distance between $\boldsymbol{s}^{(t)}$ and $\boldsymbol{z}^{(t)}$ (Equation (180)). □

Note that the bound in this theorem is not affine invariant because of the presence of Lipschitz constants and strong convexity constants of the sets. The algorithm is not affine invariant either because the step size rule depends on these constants as well as on $\|\boldsymbol{s}^{(t)} - \boldsymbol{z}^{(t)}\|$. Deriving an affine invariant step size choice and convergence analysis is still an interesting open problem in this setting.

F Details on the experiments

Graphical Games. The payoff matrix M that we use encodes the following simple model of competition between universities with their respective benefits:

1. University 1 (respectively University 2) has benefit $b_i^{(1)}$ ($b_i^{(2)}$) to get student i.
2. Student i ranks the possible roommates with a permutation $\sigma_i \in \mathcal{S}_p$. Let $\sigma_i(j)$ represents the rank of j for i (first in the list is the preferred one).
3. They go to the university that matched them with their preferred roommate, in case of equality the student chooses randomly.
4. Supposing that $\boldsymbol{x}$ encodes the roommate assignment proposed by University 1 (and $\boldsymbol{y}$ for University 2), then the expectation of the benefit of University 1 is $\boldsymbol{x}^\top M \boldsymbol{y}$, with the following definition for the payoff matrix M indexed by pairs of matched students. For the pairs (i,j) with $i<j$ and (k,l) with $k<l$ with elements in $1,\dots,s$, we have:

(a) $M_{ij,il} = \begin{cases} b_i^{(1)} & \text{if } \sigma_i(j) < \sigma_i(l) \quad \textit{i.e. student } i \textit{ preferred } j \textit{ over } l \\ -b_i^{(2)} & \text{if } \sigma_i(j) > \sigma_i(l) \\ \frac{b_i^{(1)} - b_i^{(2)}}{2} & \text{otherwise} \quad (\text{in that case } j = l). \end{cases}$

(b) $M_{ij,kj} = M_{ji,jk}$

(c) $M_{ij,ki} = M_{ij,ik}$

(d) $M_{ij,jl} = M_{ij,lj}$

(e) $M_{ij,kl} = 0$ otherwise

Note that we need to do unipartite matching here (and not bipartite matching) since we have to match students together and not students with dorms.

For our experiments, in order to get a realistic payoff matrix, we set $\mu_i \sim \mathcal{U}[0,1]$ the *true* value of student i. Then we set $b_i^{(U)} \sim \mathcal{N}(\mu_i, 0.1)$ the value of the student i *observed* by University U. To solve the perfect matching problem, we used Blossom V by Kolmogorov (2009).

Sparse structured SVM. We give here more details on the derivations of the objective function for the structured SVM problem. We first recall the structured prediction setup with the same notation from (Lacoste-Julien et al., 2013). In structured prediction, the goal is to predict a structured object $\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})$ (such as a sequence of tags) for a given input $\boldsymbol{x} \in \mathcal{X}$. For the structured SVM approach, a structured feature map $\boldsymbol{\phi} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ encodes the relevant information for input / output pairs, and a linear classifier with parameter $\boldsymbol{w}$ is defined by $h_{\boldsymbol{w}} = \arg\max_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})} \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) \rangle$. We are also given a task-dependent structured error $L(\boldsymbol{y}', \boldsymbol{y})$ that gives the loss of predicting $\boldsymbol{y}$ when the ground truth is $\boldsymbol{y}'$. Given a labeled training set $\{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})\}_{i=1}^n$, the standard ℓ_2-regularized structured SVM objective in its non-smooth formulation for learning as given for example in Equation (3) from (Lacoste-Julien et al., 2013) is:

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \frac{\lambda}{2} \|\boldsymbol{w}\|_2^2 + \frac{1}{n} \sum_i \tilde{H}_i(\boldsymbol{w}) \tag{188}$$

where $\tilde{H}_i(\boldsymbol{w}) := \max_{\boldsymbol{y} \in \mathcal{Y}_i} \; L_i(\boldsymbol{y}) - \langle \boldsymbol{w}, \boldsymbol{\psi}_i(\boldsymbol{y}) \rangle$ is the structured hinge loss, and the following notational shorthands were defined: $\mathcal{Y}_i := \mathcal{Y}(\boldsymbol{x}^{(i)})$, $L_i(\boldsymbol{y}) := L(\boldsymbol{y}^{(i)}, \boldsymbol{y})$ and $\boldsymbol{\psi}_i(\boldsymbol{y}) := \boldsymbol{\phi}(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) - \boldsymbol{\phi}(\boldsymbol{x}^{(i)}, \boldsymbol{y})$.

In our setting, we consider a sparsity inducing ℓ_1-regularization instead. Moreover, we use the (equivalent) constrained formulation instead of the penalized one, in order to get a problem over a polytope. We thus get the following challenging problem:

$$\min_{\|\boldsymbol{w}\|_1 \leq R} \frac{1}{n} \sum_i \tilde{H}_i(\boldsymbol{w}). \tag{189}$$

To handle any type of structured output space $\mathcal{Y}$, we use the following generic encoding. Enumerating the elements of $\mathcal{Y}_i$, we can represent the j^{th} element of $\mathcal{Y}_i$ as $(\overbrace{0, \dots, 0}^{j-1}, 1, 0, \dots, 0) \in \mathbb{R}^{|\mathcal{Y}_i|}$. Let M_i have $\big(\boldsymbol{\psi}_i(\boldsymbol{y})\big)_{\boldsymbol{y} \in \mathcal{Y}_i}$ as columns and let $\boldsymbol{L}_i$ be a vector of length $|\mathcal{Y}_i|$ with $L_i(\boldsymbol{y})$ as its entries. The functions $\tilde{H}_i(\boldsymbol{w})$ can then be rewritten as the maximization of linear functions in $\boldsymbol{y}$: $\tilde{H}_i(\boldsymbol{w}) = \max_{\boldsymbol{y} \in \mathcal{Y}_i} \; \boldsymbol{L}_i^\top \boldsymbol{y} - \boldsymbol{w}^\top M_i \boldsymbol{y}$. As the maximization of linear functions over a polytope is always obtained at one of its vertex, we can equivalently define the maximization over the convex hull of $\mathcal{Y}_i$, which is the probability simplex in $\mathbb{R}^{|\mathcal{Y}_i|}$ that we denote $\Delta(|\mathcal{Y}_i|)$:

$$\max_{\boldsymbol{y}_i \in \mathcal{Y}_i} \; \boldsymbol{L}_i^\top \boldsymbol{y}_i - \boldsymbol{w}^\top M_i \boldsymbol{y}_i = \max_{\boldsymbol{\alpha}_i \in \Delta(|\mathcal{Y}_i|)} \boldsymbol{L}_i^\top \boldsymbol{\alpha}_i - \boldsymbol{w}^\top M_i \boldsymbol{\alpha}_i \tag{190}$$

Thus our equivalent objective is

$$\min_{\|\boldsymbol{w}\|_1 \leq R} \frac{1}{n} \sum_i \Big(\max_{\boldsymbol{y}_i \in \mathcal{Y}_i} \; \boldsymbol{L}_i^\top \boldsymbol{y}_i - \boldsymbol{w}^\top M_i \boldsymbol{y}_i \Big) = \min_{\|\boldsymbol{w}\|_1 \leq R} \frac{1}{n} \sum_i \Big(\max_{\boldsymbol{\alpha}_i \in \Delta(|\mathcal{Y}_i|)} \boldsymbol{L}_i^\top \boldsymbol{\alpha}_i - \boldsymbol{w}^\top M_i \boldsymbol{\alpha}_i \Big), \tag{191}$$

which is the bilinear saddle point formulation given in the main text in (31).

Supplementary References

V. Kolmogorov. Blossom V: a new implementation of a minimum cost perfect matching algorithm. *Mathematical Programming Computation*, 2009.

S. Lacoste-Julien and M. Jaggi. An affine invariant linear convergence analysis for Frank-Wolfe algorithms. *arXiv preprint arXiv:1312.7864*, 2013.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer, 2004.

A. Osokin, J.-B. Alayrac, I. Lukasewitz, P. K. Dokania, and S. Lacoste-Julien. Minding the gaps for block Frank-Wolfe optimization of structured SVMs. In *ICML*, 2016.