Appendix

A Primal-Dual Structure

Apart from the notations discussed in the main paper, we would further use the following notation for data matrix $X \in \mathbb{R}^{n\times d}$ such that $X = \left[x_1^\top; \cdots; x_n^\top\right]$. We consider the following general primal and its corresponding dual problem which appear very frequently in machine learning domain.

$$\min_{\theta\in\mathbb{R}^d} \left[\mathcal{O}_P(\theta) := \psi(\theta) + \frac{1}{n}\sum_{i=1}^n \phi_i(x_i^\top\theta)\right] \tag{15}$$

$$\max_{\alpha\in\mathbb{R}^n} \left[\mathcal{O}_D(\alpha) := -\psi^*\left(-\frac{1}{n}X^\top\alpha\right) - \frac{1}{n}\sum_{i=1}^n \phi_i^*(\alpha_i)\right]. \tag{16}$$

Here, we assume that $\phi : \mathbb{R}^d \to \mathbb{R}$ and $\psi : \mathbb{R}^d \to \mathbb{R}$ are smooth convex function for all i. We have the following first order optimality conditions for the equivalent problems given in Equations (15) and (16):

$$\begin{aligned} &x_i^\top\theta \in \partial\phi_i^*(\alpha_i), \\ &\theta \in \partial\psi^*\left(-\frac{1}{n}X^\top\alpha\right), \end{aligned} \quad \text{and} \quad \begin{aligned} &\alpha_i \in \partial\phi_i(x_i^\top\theta), \\ &-\frac{1}{n}X^\top\alpha \in \partial\psi(\theta). \end{aligned} \tag{17}$$

From the duality, $\theta(\alpha) = \partial\psi^*\left(-\frac{1}{n}\sum_{i=1}^n \alpha_i x_i\right)$. We can recall Fenchel's Inequality: For any convex function f, the inequality $f(x) + f^*(\theta) \geq x^\top\theta$ holds for all $x \in \mathrm{dom}(f)$ and $\theta \in \mathrm{dom}(f^*)$. Equality holds if the following is satisfied $\theta \in \partial f(x)$.

From Fenchel's inequality, we have:

Proposition 2 *Consider the general primal dual problem given in equations* (15) *and* (16)*, dual sub-optimlaity gap* $gap(\alpha) = [\mathcal{O}_D(\alpha^\star) - \mathcal{O}_D(\alpha)]$ *at some* α *provides the upper bound on the Bregman divergence of* ψ *between* $\theta^\star$ *and* $\theta(\alpha)$ *i.e.* $D_\Psi(\theta^\star, \theta(\alpha)) \leqslant \mathrm{gap}(\alpha)$.

Proof The Bregman divergence with respect to mirror map ψ is

$$D_\Psi(x, y) = \psi(x) - \psi(y) - \langle\nabla\psi(y), x - y\rangle.$$

Now, we have:

$$\mathrm{gap}(\alpha) = -\psi^*\left(-\frac{1}{n}X^\top\alpha^\star\right) + \psi^*\left(-\frac{1}{n}X^\top\alpha\right) - \frac{1}{n}\sum_{i=1}^n \phi_i^*(\alpha_i^\star) + \frac{1}{n}\sum_{i=1}^n \phi_i^*(\alpha_i). \tag{18}$$

In the proof we would again use Fenchel's inequality which we used in the proof of previous theorem. From the optimality condition, we know that $-\frac{1}{n}X^\top\alpha \in \partial\psi(\theta(\alpha))$. Hence,

Hence,

$$\begin{aligned} \mathrm{gap}(\alpha) &= -\psi^*\left(-\frac{1}{n}X^\top\alpha^\star\right) + \psi^*\left(-\frac{1}{n}X^\top\alpha\right) - \frac{1}{n}\sum_{i=1}^n \phi_i^*(\alpha_i^\star) + \frac{1}{n}\sum_{i=1}^n \phi_i^*(\alpha_i) \\ &= -\left(-\left\langle\frac{1}{n}X^\top\alpha^\star, \theta^\star\right\rangle - \psi(\theta^\star)\right) + \left(-\left\langle\frac{1}{n}X^\top\alpha, \theta(\alpha)\right\rangle - \psi(\theta(\alpha))\right) - \frac{1}{n}\sum_{i=1}^n \phi_i^*(\alpha_i^\star) + \frac{1}{n}\sum_{i=1}^n \phi_i^*(\alpha_i) \\ &= \psi(\theta^\star) - \psi(\theta(\alpha)) + \left\langle\frac{1}{n}X^\top\alpha^\star, \theta^\star\right\rangle - \left\langle\frac{1}{n}X^\top\alpha, \theta(\alpha)\right\rangle - \frac{1}{n}\sum_{i=1}^n \phi_i^*(\alpha_i^\star) + \frac{1}{n}\sum_{i=1}^n \phi_i^*(\alpha_i) \\ &= \psi(\theta^\star) - \psi(\theta(\alpha)) + \left\langle\frac{1}{n}X^\top\alpha^\star, \theta^\star\right\rangle - \left\langle\frac{1}{n}X^\top\alpha, \theta(\alpha)\right\rangle - \frac{1}{n}\sum_{i=1}^n \phi_i^*(\alpha_i^\star) + \frac{1}{n}\sum_{i=1}^n \phi_i^*(\alpha_i) \end{aligned}$$

$$
\begin{aligned}
&= \psi(\theta^\star) - \psi(\theta(\alpha)) + \left\langle \frac{1}{n} X^\top \alpha^\star, \theta^\star \right\rangle + \left\langle \frac{1}{n} X^\top \alpha, \theta^\star \right\rangle - \left\langle \frac{1}{n} X^\top \alpha, \theta^\star \right\rangle - \left\langle \frac{1}{n} X^\top \alpha, \theta(\alpha) \right\rangle \\
&\qquad\qquad - \frac{1}{n}\sum_{i=1}^n \phi_i^*(\alpha_i^\star) + \frac{1}{n}\sum_{i=1}^n \phi_i^*(\alpha_i) \\
&= \psi(\theta^\star) - \psi(\theta(\alpha)) - \left\langle \frac{1}{n} X^\top \alpha, \theta(\alpha) - \theta^\star \right\rangle + \left\langle \frac{1}{n} X^\top \alpha^\star - \frac{1}{n} X^\top \alpha, \theta^\star \right\rangle - \frac{1}{n}\sum_{i=1}^n \phi_i^*(\alpha_i^\star) + \frac{1}{n}\sum_{i=1}^n \phi_i^*(\alpha_i) \\
&= \underbrace{\psi(\theta^\star) - \psi(\theta(\alpha)) - \langle \nabla\psi(\theta(\alpha)), \theta^\star - \theta(\alpha) \rangle}_{:= D_\Psi(\theta^\star, \theta(\alpha))} + \left\langle \frac{1}{n} X^\top \alpha^\star - \frac{1}{n} X^\top \alpha, \theta^\star \right\rangle \\
&\qquad\qquad - \frac{1}{n}\sum_{i=1}^n \phi_i^*(\alpha_i^\star) + \frac{1}{n}\sum_{i=1}^n \phi_i^*(\alpha_i) \\
&= D_\Psi(\theta^\star, \theta(\alpha)) + \left\langle \frac{1}{n}\alpha^\star - \frac{1}{n}\alpha, X\theta^\star \right\rangle - \frac{1}{n}\sum_{i=1}^n \phi_i^*(\alpha_i^\star) + \frac{1}{n}\sum_{i=1}^n \phi_i^*(\alpha_i) \\
&= D_\Psi(\theta^\star, \theta(\alpha)) + \frac{1}{n}\sum_{i=1}^n (\alpha_i^\star - \alpha_i) \cdot x_i^\top \theta^\star - \frac{1}{n}\sum_{i=1}^n \phi_i^*(\alpha_i^\star) + \frac{1}{n}\sum_{i=1}^n \phi_i^*(\alpha_i) \\
&= D_\Psi(\theta^\star, \theta(\alpha)) - \frac{1}{n}\sum_{i=1}^n (\alpha_i - \alpha_i^\star) \cdot \nabla\phi^*(\alpha_i^\star) - \frac{1}{n}\sum_{i=1}^n \phi_i^*(\alpha_i^\star) + \frac{1}{n}\sum_{i=1}^n \phi_i^*(\alpha_i) \\
&= D_\Psi(\theta^\star, \theta(\alpha)) + \frac{1}{n}\sum_{i=1}^n D_{\phi_i^*}(\alpha_i, \alpha_i^\star) \geq D_\Psi(\theta^\star, \theta(\alpha)). \qquad (19)
\end{aligned}
$$

■

After we provide the general result in Proposition 2, we now provide the proof for proposition 1 below. The result in statement is a useful result and can be useful in several ways. For example, the guarantees for SDCA (Shalev-Shwartz & Zhang, 2013, 2014). We provide the details in the Appendix C.

Proof [Proof of Proposition 1] We can just use the result in Proposition 2 to prove Proposition 1. Let's recall once again the primal dual formulation of the problem which we have in Equation (3) and Equation (4).

$$
\begin{aligned}
&\min_{\theta\in\mathbb{R}^d} D_\psi(\theta, \theta^{(0)}) \text{ such that } \forall i \in \{1, \ldots, n\},\ x_i^\top \theta \in \mathcal{Y}_i \qquad (20) \\
&= \min_{\theta\in\mathbb{R}^d} \psi(\theta) + \frac{1}{n}\sum_{i=1}^n \max_{\alpha_i\in\mathbb{R}^k} \left\{ \alpha_i^\top x_i^\top \theta - \sigma_{\mathcal{Y}_i}(\alpha_i) \right\} \\
&= \max_{\forall i,\ \alpha_i\in\mathbb{R}^k} -\frac{1}{n}\sum_{i=1}^n \sigma_{\mathcal{Y}_i}(\alpha_i) - \psi^\star\Big(-\frac{1}{n}\sum_{i=1}^n x_i\alpha_i \Big) \qquad (21) \\
&= \max_{\alpha\in\mathbb{R}^{n\times k}} G(\alpha),
\end{aligned}
$$

Let $\mathcal{K}_i$ represents that set for all θ such that $x_i^\top\theta \in \mathcal{Y}_i$ and the indicator function $\iota_{\mathcal{K}_i}$ for a convex set $\mathcal{K}_i$ for all $\in \{1, \ldots, n\}$ is defined as $\iota_{\mathcal{K}_i}(x_i^\top\theta) = 0$ if $x_i^\top\theta \in \mathcal{Y}_i$ and $\iota_{\mathcal{K}_i}(x_i^\top\theta) = +\infty$, otherwise for all $\in \{1, \ldots, n\}$. We can write Equation (20) in the form of generalized equation given in Equation (15) considering $\phi_i(x_i^\top\theta) = \iota_{\mathcal{K}_i}(x_i^\top\theta)$. It is easy to see that $\phi_i^*(\alpha_i) = \sigma_{\mathcal{Y}_i}(\alpha_i)$. Hence, now the statement follows from Proposition 2. ■

A.1 Coordinate Descent Update: Proof of Lemma 1

We have:

$$
\alpha_{i(t)}^{(t)} = \arg\max_{\alpha_{i(t)}} -\frac{1}{n}\sigma_{\mathcal{Y}_{i(t)}}(\alpha_i) + \frac{1}{n}\nabla\psi^\star\Big(-\frac{1}{n}\sum_{i=1}^n x_i\alpha_i^{(t-1)} \Big)^\top x_{i(t)}[\alpha_{i(t)} - \alpha_{i(t)}^{(t-1)}] - \frac{L_{i(t)}}{2n^2}\|\alpha_i - \alpha_{i(t)}^{(t-1)}\|_2^2
$$

$$= \arg\max_{\alpha_{i(t)}} -\frac{1}{n}\sigma_{\mathcal{Y}_{i(t)}}(\alpha_i) + \frac{1}{n}\theta(\alpha^{(t-1)})^\top x_{i(t)}[\alpha_{i(t)} - \alpha_{i(t)}^{(t-1)}] - \frac{L_{i(t)}}{2n^2}\|\alpha_i - \alpha_{i(t)}^{(t-1)}\|_2^2$$
$$= \arg\min_{\alpha_{i(t)}} \sigma_{\mathcal{Y}_{i(t)}}(\alpha_i) + \frac{L_{i(t)}}{2n}\|\alpha_i - \alpha_{i(t)}^{(t-1)} - \frac{n}{L_{i(t)}}x_{i(t)}^\top\theta(\alpha^{(t-1)})\|_2^2. \tag{22}$$

The minimization problem in Equation (22) can be written as follows:

$$\begin{aligned}
&\min_{\alpha_{i(t)}}\left[\sigma_{\mathcal{Y}_{i(t)}}(\alpha_i) + \frac{L_{i(t)}}{2n}\|\alpha_i - \alpha_{i(t)}^{(t-1)} - \frac{n}{L_{i(t)}}x_{i(t)}^\top\theta(\alpha^{(t-1)})\|_2^2\right]\\
=&\min_{\alpha_{i(t)}}\left[\sigma_{\mathcal{Y}_{i(t)}}(\alpha_i) - \sup_z\left[(\alpha_i - \alpha_{i(t)}^{(t-1)} - \frac{n}{L_{i(t)}}x_{i(t)}^\top\theta(\alpha^{(t-1)}))^\top z + \frac{n}{2L_{i(t)}}\|z\|^2\right]\right]\\
=&\sup_{z\in\mathcal{Y}_{i(t)}}\left[-\frac{n}{2L_{i(t)}}\|z\|^2 + z^\top\left(\frac{n}{L_{i(t)}}x_{i(t)}^\top\theta(\alpha^{(t-1)}) + \alpha_{i(t)}^{(t-1)}\right)\right]
\end{aligned} \tag{23}$$

The above maximization problem has a solution at $z^\star = \Pi_{\mathcal{Y}_{i(t)}}\left(x_{i(t)}^\top\theta(\alpha^{(t-1)}) + \frac{L_{i(t)}}{n}\alpha_{i(t)}^{(t-1)}\right)$. However, $z^\star$ is also the solution of the following optimization formulation:

$$z^\star = \arg\max_z\left[(\alpha_i - \alpha_{i(t)}^{(t-1)} - \frac{n}{L_{i(t)}}x_{i(t)}^\top\theta(\alpha^{(t-1)}))^\top z + \frac{n}{2L_{i(t)}}\|z\|^2\right]$$

Comparing both the value of $z^\star$, we get the following update in $\alpha_{i(t)}$ in alternative form

$$\alpha_{i(t)} = \alpha_{i(t)}^{(t-1)} + \frac{n}{L_{i(t)}}x_{i(t)}^\top\theta(\alpha^{(t-1)}) - \frac{n}{L_{i(t)}}\Pi_{\mathcal{Y}_i}\Big(\frac{L_{i(t)}}{n}\alpha_{i(t)}^{(t-1)} + x_{i(t)}^\top\theta(\alpha^{(t-1)})\Big),$$

where $\Pi_{\mathcal{Y}_i}$ is the orthogonal projection on $\mathcal{Y}_i$.

A.2 Mirror Descent: [Proof of Theorem 1]

The convergence rate does depend on $\psi(\theta^\star)$ but this is not an explicit regularization. The proof goes as follows: Mirror descent with the mirror map ψ selects $i(t)$ at random and the iteration is

$$\psi'(\theta^{(t)}) = \psi'(\theta^{(t-1)}) - \gamma x_{i(t)}(\Pi_{\mathcal{Y}_i}(x_{i(t)}^\top\theta^{(t-1)}) - x_{i(t)}^\top\theta^{(t-1)}).$$

Following the proof of Flammarion & Bach (2017), we have for any $\theta \in \mathbb{R}^d$:

$$\begin{aligned}
D_\psi(\theta, \theta^{(t)}) &= D_\psi(\theta, \theta^{(t)}) - D_\psi(\theta^{(t)}, \theta^{(t-1)}) + \gamma f_t'(\theta^{(t-1)})^\top(\theta^{(t)} - \theta)\\
&\leqslant D_\psi(\theta, \theta^{(t)}) - \frac{\mu}{2}\|\theta^{(t)} - \theta^{(t-1)}\|^2 + \gamma f_t'(\theta^{(t-1)})^\top(\theta^{(t-1)} - \theta)\\
&\quad +\gamma\|f_t'(\theta^{(t-1)})\|_\star\|\theta^{(t-1)} - \theta^{(t)}\|\\
&\leqslant D_\psi(\theta, \theta^{(t)}) - \gamma f_t'(\theta^{(t-1)})^\top(\theta^{(t-1)} - \theta) + \frac{\gamma^2}{2\mu}\|f_t'(\theta^{(t-1)})\|_\star^2.
\end{aligned}$$

For $\theta = \theta^\star$ and using $\mathbb{E}\big[\|f_t'(\theta^{(t-1)})\|_\star^2\big] \leqslant \sup_i \|x_i\|_{2\to\star}^2\big[f(\theta) - f(\theta^\star)\big]$, we get and taking expectations, we get:

$$\big(1 - \gamma\frac{\|x_i\|_{2\to\star}^2}{2\mu}\big)\mathbb{E}\big[f(\theta^{(t-1)}) - f(\theta^\star)\big] \leqslant \frac{1}{\gamma}\Big(\mathbb{E}[D_\psi(\theta^\star, \theta^{(t)})] - \mathbb{E}\big[D_\psi(\theta^\star, \theta^{(t-1)})\big]\Big).$$

Thus, with $\gamma = \mu/\sup_i \|x_i\|_{2\to\star}^2$, we get

$$\mathbb{E}\big[f(\theta^{(t-1)}) - f(\theta^\star)\big] \leqslant \frac{2}{\gamma}\Big(\mathbb{E}[D_\psi(\theta^\star, \theta^{(t)})] - \mathbb{E}\big[D_\psi(\theta^\star, \theta^{(t-1)})\big]\Big).$$

This leads to

$$\mathbb{E}\big[f(\bar{\theta}_t) - f(\theta^\star)\big] \leqslant \frac{2}{\gamma t}D_\psi(\theta^\star, \theta^{(0)}).$$

B ℓ_p-perceptron

In this section, we provide proofs for the claims made in Section 3.

We start with the proof of Lemma 3.

Proof For all i, $x_i^\top \theta^\star \geq 1$. Hence,

$$\begin{aligned} x_i^\top \theta_t &= x_i^\top \theta_t - x_i^\top \theta^\star + x_i^\top \theta^\star = x_i^\top \theta^\star - x_i^\top (\theta^\star - \theta_t) \\ &\geq 1 - x_i^\top (\theta^\star - \theta_t) \geq 1 - \|x_i\|_q \|\theta_t - \theta^\star\|_p \\ &\geq 1 - R\|\theta_t - \theta^\star\|_p. \end{aligned}$$

Assuming $\alpha_0 = 0$, from Equation (13), we have

$$\mathbb{E}\Big[\|\theta_t - \theta^\star\|_p\Big] \leq \frac{2\sqrt{2}\max_i \|x_i\|_q}{\sqrt{(p-1)t}} \sqrt{\frac{G(\alpha^\star) - G(0)}{\max_i \|x_i\|_q} + \frac{1}{2}\|\alpha^\star\|^2}$$

Now for on average for no mis-classification for all $i \in \{1, \cdots, n\}$,

$$1 \geq R\mathbb{E}\Big[\|\theta_t - \theta^\star\|_p\Big] \Rightarrow t \geq \frac{2\sqrt{2}R^2}{\sqrt{p-1}} \sqrt{\frac{G(\alpha^\star) - G(0)}{R} + \frac{1}{2}\|\alpha^\star\|^2}. \tag{24}$$

■

Mistake Bound ℓ_p-primal perceptron. If we apply mirror descent with the mirror map $\psi = \frac{1}{2}\|\cdot\|_p^2$ to the minimization of $\frac{1}{n}\sum_{i=1}^n (1 - \theta^\top x_i)_+$, then the iteration is

$$\psi'(\theta_t) = \psi'(\theta_{t-1}) - \gamma 1_{1 - \theta_{t-1}^\top x_{i(t)} > 0} x_{i(t)},$$

and we have

$$\frac{1}{n}\sum_{i=1}^n (1 - \bar{\theta}_t^\top x_i)_+ \leqslant \frac{\|\theta_\star\|_p^2}{2\gamma t} + \gamma \frac{\max_i \|x_i\|_q^2}{2(p-1)}.$$

The best γ is equal to $\gamma = \frac{\|\theta_\star\|_p}{\max_i \|x_i\|_q} \frac{\sqrt{p-1}}{\sqrt{t}}$, which does depend on too many things, and leads to a proportion of mistakes on the training set less than

$$\frac{\|\theta_\star\|_p \max_i \|x_i\|_q}{\sqrt{p-1}\sqrt{t}}.$$

B.1 Update for Random Coordinate Descent

We have:

$$\begin{aligned} & \min_{\theta \in \mathbb{R}^d} \frac{1}{2}\|\theta\|_p^2 \text{ such that } X\theta \geqslant 1 \\ = \; & \min_{\theta \in \mathbb{R}^d} \max_{\alpha \in \mathbb{R}^n} \frac{1}{2}\|\theta\|_p^2 + \alpha^\top (1 - X\theta) \\ = \; & \max_{\alpha \in \mathbb{R}^n} -\frac{1}{2}\|X^\top \alpha\|_q^2 + \alpha^\top 1, \end{aligned}$$

where, at optimality, θ can be obtained from $X^\top \alpha$ as

$$\theta_j = \|X^\top \alpha\|_q^{2-q} (X^\top \alpha)_j^{q-1},$$

where we define $u^{q-1} = |u|^{q-1}\text{sign}(u)$.

The function $\frac{1}{2}\|X^\top \alpha\|_p^2$ is smooth, and the regular smoothness constant with respect to the i-th variable which is less than

$$L_i = \frac{1}{p-1}\|x_i\|_q^2.$$

A dual coordinate ascent step corresponds to choosing $i(t)$ and replacing $(\alpha_{t-1})_{i(t)}$ by

$$(\alpha_t)_i = \max\Big\{0, (\alpha_{t-1})_{i(t)} + \frac{1}{L_{i(t)}}\big(1 - \|X^\top \alpha_{t-1}\|_q^{2-q} \sum_{j=1}^{d} [(X^\top \alpha_{t-1})_j]^{q-1} X_{i(t)j}\Big\},$$

which can be interpreted as:

$$(\alpha_t)_i = \max\Big\{0, (\alpha_{t-1})_{i(t)} + \frac{1}{L_{i(t)}}\big(1 - \theta_{t-1}^\top x_{i(t)}\big)\Big\}.$$

B.2 ℓ_2-perceptron

The primal problem has the following dual form under the interpolation regime

$$\max_{\alpha \geq 0, \alpha \in \mathbb{R}^n} \ \alpha^\top 1 - \frac{1}{2}\|X\alpha\|^2.$$

We denote S_v as the set of support vectors *i.e.* S_v is the set of indices where $\alpha_j^\star \neq 0$. Hence, we also have $\tilde{x}_j^\top \theta^\star = 1$ for $j \in S_v$. α_{S_v} denotes the vector of non-zero entries in α. Correspondingly, X_{S_v} denotes the feature matrix for support vectors. From the first order suboptimality condition we have,

$$\theta(\alpha) = \frac{1}{n} X\alpha.$$

We also know that for support vectors, $y_i \cdot x_i^\top \theta^\star = \tilde{x}_i^\top \theta^\star = 1$ for all $i \in S_v$. Also $\theta^\star = \frac{1}{n} X_{S_v} \alpha_{S_v}^\star$. Hence,

$$\frac{1}{n} X_{S_v}^\top X_{S_v} \alpha_{S_v}^\star = 1 \Rightarrow \alpha_{S_v}^\star = n(X_{S_v}^\top X_{S_v})^{-1} 1.$$

From Lemma 3, we should have $t \geq \frac{2\sqrt{2}R^2}{\sqrt{p-1}} \sqrt{\frac{G(\alpha^\star) - G(0)}{R} + \frac{1}{2}\|\alpha^\star\|^2}$, for no training mistakes.

We now use Corollary 2 to get mistake bound on the perceptron. To have no mistakes on average, the proportion of mistakes should be less than $1/n$. Hence,

$$\frac{R\|\theta^\star\|}{\sqrt{t}} \leq \frac{1}{n} \Rightarrow t \geq R^2 \|\theta^\star\|^2 n^2. \tag{25}$$

We already have $\alpha_{S_v}^\star = n(X_{S_v}^\top X_{S_v})^{-1} 1$.

$$\theta^\star = \frac{1}{n} X\alpha^\star = \frac{1}{n} X_{S_v} \alpha_{S_v}^\star = X_{S_v}(X_{S_v}^\top X_{S_v})^{-1} 1.$$

Finally we have the following:

$$\begin{aligned} \|\alpha^\star\| &= \|\alpha_{S_v}^\star\| = n\|(X_{S_v}^\top X_{S_v})^{-1} 1\| \\ \|\theta^\star\|^2 &= \|X_{S_v}(X_{S_v}^\top X_{S_v})^{-1} 1\|^2 = 1^\top (X_{S_v}^\top X_{S_v})^{-1} 1. \end{aligned} \tag{26}$$

Hence, one can compare the number of minimum iteration required by both the approaches.

C (Accelerated) Stochastic Dual Coordinate Descent

Stochastic dual coordinate ascent (Shalev-Shwartz & Zhang, 2013) is a popular approach to optimize regularized empirical risk minimize problem. For this section, let $\phi_1, \cdots, \phi_n$ be a sequence of $\frac{1}{\gamma}$-smooth convex losses and let $\lambda > 0$ be a regularization parameter then consider following regularized empirical risk minimization problem:

$$\min_{\theta \in \mathbb{R}} \left[\mathcal{S}_P(\theta) := \frac{\lambda}{2}\|\theta\|^2 + \frac{1}{n}\sum_{i=1}^{n} \phi_i(X_i^\top \theta) \right]. \tag{27}$$

Corresponding dual problem of the minimization problem given in equation (27) can be written similarly as:

$$\max_{\alpha \in \mathbb{R}^n} \left[\mathcal{S}_D(\alpha) := -\frac{\lambda}{2} \| \frac{1}{\lambda n} X^\top \alpha \|^2 - \frac{1}{n} \sum_{i=1}^n \phi_i^*(-\alpha_i) \right] \tag{28}$$

There is one to one relation between the smoothness constant and strong convexity parameter of primal and corresponding dual function. We prove the following result from Kakade et al. (2009).

Theorem 2 (Theorem 6, (Kakade et al., 2009)) *Assume that f is a closed and convex function. Then f is β-strongly convex w.r.t. a norm $\|\cdot\|$ if and only if f^* is $\frac{1}{\beta}$-smooth w.r.t. the dual norm $\|\cdot\|_*$.*

From the above theorem it is clear that ϕ_i^* are γ-strongly convex. Hence the term $\frac{1}{n}\sum_{i=1}^n \phi_i^*(-\alpha_i)$ is $\frac{\gamma}{n}$ strongly convex. Similary coordinate wise smoothness $L_i = \frac{\|x_i\|^2}{\lambda n^2}$.
Now, just as a direct implication of the result provided in Proposition 2, we have the convergence result for SDCA (Shalev-Shwartz & Zhang, 2013) and accelerated stochastic dual coordinate ascent (Shalev-Shwartz & Zhang, 2014) which we provide in Corollary C.1 and Corollary C.2. For the next two results, we denote θ_k as $\theta(\alpha_k)$.

Corollary C.1 (Stochastic Dual Coordinate Ascent) *Consider the regularized empirical risk minimization problem given in equation (27), then if we run SDCA (Shalev-Shwartz & Zhang, 2013) algorithm starting from $\alpha_0 \in \mathbb{R}^n$ with a fix step size $1/\max_i L_i$ where $L_i = \frac{\|x_i\|^2}{\lambda n^2}$, primal iterate after k iterations converges as following:*

$$\frac{\lambda}{2}\|\theta_{k+1} - \theta^\star\|^2 \leq D(\alpha_{k+1}) \leq \left(1 - \frac{\gamma\lambda}{\max_i \|x_i\|^2}\right)^k (\mathcal{S}_D(\alpha_0) - \mathcal{S}_D(\alpha^\star)).$$

Proof From Allen-Zhu et al. (2016), it is clear that for μ-strongly convex and L_i-coordinate wise smooth convex function $\mathcal{S}_D(\alpha)$ where $\alpha \in \mathbb{R}^n$, randomized coordinate descent has the following convergence guarantee:

$$D(\alpha_{k+1}) \leq \left(1 - \frac{\mu}{n \max_i L_i}\right)^k (\mathcal{S}_D(\alpha_0) - \mathcal{S}_D(\alpha^\star)).$$

Here, $\mu = \frac{\gamma}{n}$. First part of the inequality directly comes from Proposition 2 by the observation that here $\psi(\cdot) = \frac{\lambda}{2}\|\cdot\|^2$ and bregman divergence are always positive. ■

Corollary C.2 (Accelerated Stochastic Dual Coordinate Ascent) *Consider the regularized empirical risk minimization problem given in equation (27), then if we run Accelerated SDCA (Shalev-Shwartz & Zhang, 2014) algorithm starting from $\alpha_0 \in \mathbb{R}^n$, we have following convergence rate for the primal iterates:*

$$\frac{\lambda}{2}\|\theta_{k+1} - \theta^\star\|^2 \leq D(\alpha_{k+1}) \leq 2\left(1 - \frac{\sqrt{\gamma\lambda}}{\sqrt{\max_i \|x_i\|^2}}\right)^k (\mathcal{S}_D(\alpha_0) - \mathcal{S}_D(\alpha^\star)).$$

Proof From Allen-Zhu et al. (2016), it is clear that for μ-strongly convex and L_i-coordinate wise smooth convex function $\mathcal{S}_D(\alpha)$ where $\alpha \in \mathbb{R}^n$, accelerated randomized coordinate descent has the following convergence guarantee:

$$D(\alpha_{k+1}) \leq 2\left(1 - \frac{\sqrt{\mu}}{n\sqrt{\max_i L_i}}\right)^k (\mathcal{S}_D(\alpha_0) - \mathcal{S}_D(\alpha^\star)).$$

First part of the inequality directly comes from Proposition 2 by the observation that here $\psi(\cdot) = \frac{\lambda}{2}\|\cdot\|^2$ and bregman divergence are always positive. Here $\mu = \frac{\gamma}{n}$ and $L_i = \frac{\|x_i\|^2}{\lambda n^2}$. ■

Discussion. Let us denote duality gap at dual variable α as $\Delta(\alpha)$. From the definition of the duality gap $\Delta(\alpha) = \mathcal{S}_P(\theta(\alpha)) - \mathcal{S}_D(\alpha)$. However, $\Delta(\alpha)$ is an upper bound on the primal sub-optimality gap as well on dual sub-optimality gap. The main difference in the analysis presented in our work with the works of Shalev-Shwartz & Zhang (2013) and Shalev-Shwartz & Zhang (2014) is that the we provide the guarantee in term of the iterate. However Shalev-Shwartz & Zhang (2013) and Shalev-Shwartz & Zhang (2014) provide convergence in terms of duality gap $\Delta(\alpha)$. Another main difference is that we use constant step size in each step and the output of our algorithm doesn't need averaging of the past iterates. Our analysis holds for the last iterate.