Lasso

1. Setup

We consider regularized empirical loss minimization problems of the following form:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \left\{ P(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^n \ell_i(\mathbf{x}_i^T\mathbf{w}) + \lambda\left(\frac{\|\mathbf{w}\|^2}{2} + \mathbf{w}^T\mathbf{b}\right)\right\} \tag{1}$$

According to my calculations this should be the correct conjugate of the dual loss we want Here the vectors $\{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^d$ represent the training data examples, $\mathbf{b} \in \mathbb{R}^d$ and the $\ell_i(.)$ are arbitrary convex real-valued loss functions (e.g., hinge loss), possibly depending on label information for the i-th datapoints. The constant $\lambda > 0$ is the regularization parameter.

Dual Problem, and Primal-Dual Certificates. The conjugate dual of (1) takes following form:

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \left\{ \mathcal{D}(\boldsymbol{\alpha}) := -\frac{1}{n}\sum_{j=1}^n \ell_j^*(-\alpha_j) - \frac{\lambda}{2}\left\|\frac{A\boldsymbol{\alpha}}{\lambda n} - \mathbf{b}\right\|^2\right\} \tag{2}$$

Here the data matrix $A = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ collects all data-points as its columns, and ℓ_j^* is the conjugate function to ℓ_j. See, e.g., (Shalev-Shwartz & Zhang, 2013) for several concrete applications.

It is possible to assign for any dual vector $\boldsymbol{\alpha} \in \mathbb{R}^n$ a corresponding primal feasible point

$$\mathbf{w}(\boldsymbol{\alpha}) = \tfrac{1}{\lambda n}A\boldsymbol{\alpha} - \mathbf{b} \tag{3}$$

The duality gap function is then given by:

$$G(\boldsymbol{\alpha}) := \mathcal{P}(\mathbf{w}(\boldsymbol{\alpha})) - \mathcal{D}(\boldsymbol{\alpha}) \tag{4}$$

By weak duality, every value $\mathcal{D}(\boldsymbol{\alpha})$ at a dual candidate $\boldsymbol{\alpha}$ provides a lower bound on every primal value $P(\mathbf{w})$. The duality gap is therefore a certificate on the approximation quality: The distance to the unknown true optimum $P(\mathbf{w}^*)$ must always lie within the duality gap, i.e., $G(\boldsymbol{\alpha}) = \mathcal{P}(\mathbf{w}) - \mathcal{D}(\boldsymbol{\alpha}) \geq \mathcal{P}(\mathbf{w}) - \mathcal{P}(\mathbf{w}^*) \geq 0$.

In large-scale machine learning settings like those considered here, the availability of such a computable measure of approximation quality is a significant benefit during training time. Practitioners using classical primal-only methods such as SGD have no means by which to accurately detect if a model has been well trained, as $P(\mathbf{w}^*)$ is unknown.

Classes of Loss-Functions. To simplify presentation, we assume that all loss functions ℓ_i are non-negative, and

$$\ell_i(0) \leq 1 \qquad \forall i \tag{5}$$

Definition 1 (L-Lipschitz continuous loss)**.** A function $\ell_i : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz continuous if $\forall a, b \in \mathbb{R}$, we have

$$|\ell_i(a) - \ell_i(b)| \leq L|a - b| \tag{6}$$

Data Partitioning. We write $\{\mathcal{P}_k\}_{k=1}^K$ for the given partition of the datapoints $[n] := \{1, 2, \ldots, n\}$ over the K worker machines. We denote the size of each part by $n_k = |\mathcal{P}_k|$. For any $k \in [K]$ and $\boldsymbol{\alpha} \in \mathbb{R}^n$ we use the notation $\boldsymbol{\alpha}_{[k]} \in \mathbb{R}^n$ for the vector

$$(\boldsymbol{\alpha}_{[k]})_i := \begin{cases} 0, & \text{if } i \notin \mathcal{P}_k, \\ \alpha_i, & \text{otherwise.} \end{cases}$$

Local Subproblems in CoCoA$^+$. We can define a data-local subproblem of the original dual optimization problem (2), which can be solved on machine k and only requires accessing data which is already available locally, i.e., datapoints with $i \in \mathcal{P}_k$. More formally, each machine k is assigned the following local subproblem, depending only on the change in the local dual variables α_i with $i \in \mathcal{P}_k$:

$$\max_{\Delta\boldsymbol{\alpha}_{[k]}\in\mathbb{R}^n} \; \mathcal{G}_k^{\sigma'}(\Delta\boldsymbol{\alpha}_{[k]}; \mathbf{w}) \tag{7}$$

where

$$\begin{aligned} \mathcal{G}_k^{\sigma'}(\Delta\boldsymbol{\alpha}_{[k]}; \mathbf{w}) := &-\frac{1}{n}\sum_{i\in\mathcal{P}_k} \ell_i^*(-\alpha_i - (\Delta\boldsymbol{\alpha}_{[k]})_i) \\ &- \frac{1}{K}\frac{\lambda}{2}\|\mathbf{w}\|^2 - \frac{1}{n}\mathbf{w}^T A\Delta\boldsymbol{\alpha}_{[k]} \\ &- \frac{\lambda}{2}\sigma'\left\|\frac{1}{\lambda n}A\Delta\boldsymbol{\alpha}_{[k]}\right\|^2 \end{aligned} \tag{8}$$

We are lucky here; this is the same as for CoCoA+. When written in terms of $\mathbf{w}$ the local dual problem should not change!

Interpretation. The above definition of the local objective functions $\mathcal{G}_k^{\sigma'}$ are such that they closely approximate the global dual objective $\mathcal{D}$, as we vary the 'local' variable $\Delta\boldsymbol{\alpha}_{[k]}$, in the following precise sense:

Lemma 2. *For any dual $\boldsymbol{\alpha}, \Delta\boldsymbol{\alpha} \in \mathbb{R}^n$, primal $\mathbf{w} = \mathbf{w}(\boldsymbol{\alpha})$ and real values γ, σ' satisfying (9), it holds that*

$$\mathcal{D}\Big(\boldsymbol{\alpha}+\gamma\sum_{k=1}^{K}\Delta\boldsymbol{\alpha}_{[k]}\Big) \ge (1-\gamma)\mathcal{D}(\boldsymbol{\alpha})+\gamma\sum_{k=1}^{K}\mathcal{G}_k^{\sigma'}(\Delta\boldsymbol{\alpha}_{[k]};\mathbf{w}).$$

Checked Lemma 2, statement stays the same and the proof almost too (the two changes are in blue in the proof)

The role of the parameter σ' is to measure the difficulty of the given data partition. For our purposes, we will see that it must be chosen not smaller than I suspect the definition of this will change a bit

$$\sigma' \ge \sigma'_{min} := \gamma \max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \frac{\|A\boldsymbol{\alpha}\|^2}{\sum_{k=1}^{K}\|A\boldsymbol{\alpha}_{[k]}\|^2} \qquad (9)$$

In the following Lemma 1, we show that this parameter can be upper-bounded by γK, which is trivial to calculate for all values $\gamma \in \mathbb{R}$. We show experimentally (Section **??**) that this safe upper bound for σ' has a minimal effect on the overall performance of the algorithm. Our main theorems below show convergence rates dependent on $\gamma \in [\frac{1}{K}, 1]$, which we refer to as the *aggregation parameter*.

lemma The choice of $\sigma' := \gamma K$ is valid for (9), i.e.,

$$\gamma K \ge \sigma'_{min}$$

Notion of Approximation Quality of the Local Solver.

Assumption 1 (Θ-approximate solution)**.** *We assume that there exists $\Theta \in [0, 1)$ such that $\forall k \in [K]$, the local solver at any iteration t produces a (possibly) randomized approximate solution $\Delta\boldsymbol{\alpha}_{[k]}$, which satisfies*

$$\mathbb{E}\big[\mathcal{G}_k^{\sigma'}(\Delta\boldsymbol{\alpha}^*_{[k]};\mathbf{w}) - \mathcal{G}_k^{\sigma'}(\Delta\boldsymbol{\alpha}_{[k]};\mathbf{w})\big] \le \qquad (10)$$
$$\Theta\left(\mathcal{G}_k^{\sigma'}(\Delta\boldsymbol{\alpha}^*_{[k]};\mathbf{w}) - \mathcal{G}_k^{\sigma'}(\mathbf{0};\mathbf{w})\right),$$

where

$$\Delta\boldsymbol{\alpha}^*_{[k]} \in \arg\max_{\Delta\boldsymbol{\alpha}\in\mathbb{R}^n} \mathcal{G}_k^{\sigma'}(\Delta\boldsymbol{\alpha}_{[k]};\mathbf{w}) \quad \forall k \in [K]. \qquad (11)$$

We are now ready to describe the CoCoA⁺ framework, shown in Algorithm 1. The crucial difference compared to the existing CoCoA algorithm (Jaggi et al., 2014) is the more general local subproblem, as defined in (8), as well as the aggregation parameter γ. These modifications allow the option of directly adding updates to the global vector $\mathbf{w}$.

2. Convergence Guarantees

Before being able to state our main convergence results, we introduce some useful quantities and the following main lemma characterizing the effect of iterations of Algorithm 1, for any chosen internal local solver.

Algorithm 1 CoCoA⁺ Framework

```
Input: Datapoints $A$ distributed according to partition $\{\mathcal{P}_k\}_{k=1}^K$. Aggregation parameter $\gamma \in [\frac{1}{K}, 1]$, subproblem parameter $\sigma'$ for the subproblems $\mathcal{G}_k^{\sigma'}(\Delta\boldsymbol{\alpha}_{[k]};\mathbf{w})$ for each $k \in [K]$.
Starting point $\boldsymbol{\alpha}^{(0)} = \mathbf{0} \in \mathbb{R}^n$, $\mathbf{w}^{(0)} := -\mathbf{b} \in \mathbb{R}^d$.
for $t = 0, 1, 2, \dots$ do
  for $k \in \{1, 2, \dots, K\}$ in parallel over computers do
    call the local solver, computing a $\Theta$-approximate solution $\Delta\boldsymbol{\alpha}_{[k]}$ of the local subproblem (8)
    update $\boldsymbol{\alpha}^{(t+1)}_{[k]} := \boldsymbol{\alpha}^{(t)}_{[k]} + \gamma\,\Delta\boldsymbol{\alpha}_{[k]}$
    return $\Delta\mathbf{w}_k := \frac{1}{\lambda n}A\Delta\boldsymbol{\alpha}_{[k]} - \mathbf{b}$
  end for
  reduce   $\mathbf{w}^{(t+1)} := \mathbf{w}^{(t)} + \gamma\sum_{k=1}^K \Delta\mathbf{w}_k$.   (12)
end for
```

Lemma 3. *Let ℓ_i^* be strongly*[1] *convex with convexity parameter $\mu \ge 0$ with respect to the norm $\|\cdot\|$, $\forall i \in [n]$. Then for all iterations t of Algorithm 1 under Assumption 1, and any $s \in [0, 1]$, it holds that*

$$\mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^{(t+1)}) - \mathcal{D}(\boldsymbol{\alpha}^{(t)})] \ge \qquad (13)$$
$$\gamma(1-\Theta)\Big(sG(\boldsymbol{\alpha}^{(t)}) - \frac{\sigma'}{2\lambda}\big(\frac{s}{n}\big)^2 R^{(t)}\Big),$$

where

$$R^{(t)} := -\tfrac{\lambda\mu n(1-s)}{\sigma' s}\|\mathbf{u}^{(t)} - \boldsymbol{\alpha}^{(t)}\|^2 \qquad (14)$$
$$+ \textstyle\sum_{k=1}^K \|A(\mathbf{u}^{(t)} - \boldsymbol{\alpha}^{(t)})_{[k]}\|^2,$$

for $\mathbf{u}^{(t)} \in \mathbb{R}^n$ with

$$-u_i^{(t)} \in \partial\ell_i(\mathbf{w}(\boldsymbol{\alpha}^{(t)})^T\mathbf{x}_i). \qquad (15)$$

Lemma 3 works exactly the same in our case, even though I'd have expected a b to show up but I don't see it from the proof. I'm not totally sure though as there are passages in the proof I don't fully understand

Provided that the Lemma 3 stays exactly the same, the next one should too. Also as it is shown in the following page, this lemma won't require any other change since I found a new surrogate to the Lasso loss (a new one, not the quadratic one) such that the conjugate is Liptschitz; we can then reuse it as it is. The following Lemma provides a uniform bound on $R^{(t)}$:

Lemma 4. *If ℓ_i are L-Lipschitz continuous for all $i \in [n]$, then*

$$\forall t : R^{(t)} \le 4L^2 \underbrace{\sum_{k=1}^K \sigma_k n_k}_{=:\sigma}, \qquad (16)$$

[1] Note that the case of weakly convex $\ell_i^*(.)$ is explicitly allowed here as well, as the Lemma holds for the case $\mu = 0$.

where

$$\sigma_k := \max_{\boldsymbol{\alpha}_{[k]}\in\mathbb{R}^n} \frac{\|A\boldsymbol{\alpha}_{[k]}\|^2}{\|\boldsymbol{\alpha}_{[k]}\|^2}. \tag{17}$$

Remark 5. *If all data-points $\mathbf{x}_i$ are normalized such that $\|\mathbf{x}_i\| \le 1 \; \forall i \in [n]$, then $\sigma_k \le |\mathcal{P}_k| = n_k$. Furthermore, if we assume that the data partition is balanced, i.e., that $n_k = n/K$ for all k, then $\sigma \le n^2/K$. This can be used to bound the constants $R^{(t)}$, above, as $R^{(t)} \le \frac{4L^2n^2}{K}$.*

2.1. Primal-Dual Convergence for General Convex Losses

I skipped over this section as here everything should follow just the same since the previous lemma didn't really change in the new setting. The following theorem shows the convergence for non-smooth loss functions, in terms of objective values as well as primal-dual gap. The analysis in (Jaggi et al., 2014) only covered the case of smooth loss functions.

Theorem 6. *Consider Algorithm 1 with $\boldsymbol{\alpha}^{(0)}$=$\mathbf{0}$$\in$$\mathbb{R}^n$. Let Assumption 1 hold, $\ell_i(\cdot)$ be L-Lipschitz continuous, and ϵ_G $>$0 be the desired duality gap (and hence an upper-bound on primal sub-optimality). Then after T iterations, where*

$$T \ge T_0 + \max\Big\{\Big\lceil \frac{1}{\gamma(1-\Theta)} \Big\rceil, \frac{4L^2\sigma\sigma'}{\lambda n^2 \epsilon_G \gamma(1-\Theta)}\Big\}, \tag{18}$$

$$T_0 \ge t_0 + \left(\frac{2}{\gamma(1-\Theta)} \left(\frac{8L^2\sigma\sigma'}{\lambda n^2 \epsilon_G} - 1 \right) \right)_+ ,$$

$$t_0 \ge \max(0, \Big\lceil \tfrac{1}{\gamma(1-\Theta)} \log(\tfrac{2\lambda n^2 (\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(0)}))}{4L^2\sigma\sigma'}) \Big\rceil),$$

we have that the expected duality gap satisfies

$$\mathbb{E}[\mathcal{P}(\mathbf{w}(\overline{\boldsymbol{\alpha}})) - \mathcal{D}(\overline{\boldsymbol{\alpha}})] \le \epsilon_G,$$

at the averaged iterate

$$\overline{\boldsymbol{\alpha}} := \tfrac{1}{T-T_0} \textstyle\sum_{t=T_0+1}^{T-1} \boldsymbol{\alpha}^{(t)}. \tag{19}$$

The following corollary of the above theorem clarifies our main result: The more aggressive adding of the partial updates, as compared averaging, offers a very significant improvement in terms of total iterations needed. While the convergence in the 'adding' case becomes independent of the number of machines K, the 'averaging' regime shows the known degradation of the rate with growing K, which is a major drawback of the original CoCoA algorithm. This important difference in the convergence speed is not a theoretical artifact but also confirmed in our practical experiments below for different K, as shown e.g. in Figure **??**.

We further demonstrate below that by choosing γ and σ' accordingly, we can still recover the original CoCoA algorithm and its rate.

Corollary 7. *Assume that all datapoints $\mathbf{x}_i$ are bounded as $\|\mathbf{x}_i\| \le 1$ and that the data partition is balanced, i.e. that $n_k = n/K$ for all k. We consider two different possible choices of the aggregation parameter γ:*

- *(CoCoA Averaging, $\gamma := \frac{1}{K}$): In this case, $\sigma' := 1$ is a valid choice which satisfies (9). Then using $\sigma \le n^2/K$ in light of Remark 5, we have that T iterations are sufficient for primal-dual accuracy ϵ_G, with*

$$T \ge T_0 + \max\Big\{\Big\lceil \frac{K}{(1-\Theta)} \Big\rceil, \frac{4L^2}{\lambda \epsilon_G (1-\Theta)}\Big\},$$

$$T_0 \ge t_0 + \left(\frac{2K}{(1-\Theta)} \left(\frac{8L^2}{\lambda K \epsilon_G} - 1 \right) \right)_+ ,$$

$$t_0 \ge \max(0, \lceil \tfrac{K}{(1-\Theta)} \log(\tfrac{2\lambda(\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(0)}))}{4KL^2}) \rceil)$$

Hence the more machines K, the more iterations are needed (in the worst case).

- *(CoCoA⁺ Adding, $\gamma := 1$): In this case, the choice of $\sigma' := K$ satisfies (9). Then using $\sigma \le n^2/K$ in light of Remark 5, we have that T iterations are sufficient for primal-dual accuracy ϵ_G, with*

$$T \ge T_0 + \max\Big\{\Big\lceil \frac{1}{(1-\Theta)} \Big\rceil, \frac{4L^2}{\lambda \epsilon_G (1-\Theta)}\Big\},$$

$$T_0 \ge t_0 + \left(\frac{2}{(1-\Theta)} \left(\frac{8L^2}{\lambda \epsilon_G} - 1 \right) \right)_+ ,$$

$$t_0 \ge \max(0, \lceil \tfrac{1}{(1-\Theta)} \log(\tfrac{2\lambda n(\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(0)}))}{4L^2K}) \rceil)$$

This is significantly better than the averaging case above.

In practice, we usually have $\sigma \ll n^2/K$, and hence the actual convergence rate can be much better than the proven worst-case bound. Table 1 shows that the actual value of σ is typically between one and two orders of magnitudes smaller compared to our used upper-bound n^2/K.

Table 1. The ratio of upper-bound $\frac{n^2}{K}$ divided by the true value of the parameter σ, for some real datasets.

K	16	32	64	128	256	512
news	15.483	14.933	14.278	13.390	12.074	10.252
real-sim	42.127	36.898	30.780	23.814	16.965	11.835
rcv1	40.138	23.827	28.204	21.792	16.339	11.099
K	256	512	1024	2048	4096	8192
covtype	17.277	17.260	17.239	16.948	17.238	12.729

2.2. Lasso

We wish to solve the following minimization problem:

$$\kappa\|\alpha\|_1 + \frac{1}{2n}\|A\alpha - b\|_2^2$$

First of all, we'll rewrite it a form compatible to the one studied in CoCoA+:

$$-\frac{1}{n}\sum_{i=1}^{n} n\kappa|\alpha_i| - \frac{\lambda}{2}\|A\alpha - b\|_2^2$$

with $\lambda = \frac{1}{n}$ and we'll define: $\ell(\alpha_i) = n\kappa|\alpha_i|$. We now notice that for an optimal solution α to this problem, it holds that for every i:

$$\ell(\alpha_i) \leq \|b\|_2^2 := B$$

since the solution with $\alpha = \mathbf{0}$ has objective value B. Therefore it must be that

$$|\alpha_i| \leq \frac{B}{n\kappa}.$$

We can therefore define a surrogate loss function $\bar{\ell}$ to be as follows:

$$\bar{\ell}(\alpha_i) = \begin{cases} n\kappa|\alpha_i| & : |\alpha_i| \leq B/(n\kappa) \\ +\infty & : otherwise \end{cases}$$

and the following problem will thus clearly have the same optimal solution as the one defined with ℓ:

$$D(\alpha) = -\frac{1}{n}\sum_{i=1}^{n} \bar{\ell}_i(\alpha_i) - \frac{\lambda}{2}\|A\alpha - b\|_2^2.$$

We'll now compute the dual conjugate of $\bar{\ell}$, which is as follows (a proof will come):

$$\bar{\ell}^*(x) = \begin{cases} 0 & : \frac{x}{n\kappa} \in [-1, 1] \\ B(|\frac{x}{n\kappa}| - 1) & : otherwise \end{cases}$$

The convenient thing about this conjugate with respect to the indicator function (the conjugate of ℓ) is of course that is defined on the entire $\mathbb{R}$. Even better, this conjugate is $B/n\kappa$ Liptschitz!

References

Jaggi, M., Smith, V., Takáč, M., Terhorst, J., Krishnan, S., Hofmann, T., and Jordan, M. I. Communication-efficient distributed dual coordinate ascent. In *NIPS*, 2014.

Richtárik, P. and Takáč, M. Distributed coordinate descent method for learning with big data. *arXiv preprint arXiv:1310.2059*, 2013.

Shalev-Shwartz, S. and Zhang, T. Stochastic Dual Coordinate Ascent Methods for Regularized Loss Minimization. *JMLR*, 14:567–599, February 2013.

Part I

Appendix

A. Technical Lemmas

Lemma 8 (Lemma 21 in (Shalev-Shwartz & Zhang, 2013))**.** *Let $\ell_i : \mathbb{R} \to \mathbb{R}$ be an L-Lipschitz continuous. Then for any real value a with $|a| > L$ we have that $\ell_i^*(a) = \infty$.*

Lemma 9. *Assuming the loss functions ℓ_i are bounded by $\ell_i(0) \leq 1$ for all $i \in [n]$ (as we have assumed in (5) above), then for the zero vector $\boldsymbol{\alpha}^{(0)} := \mathbf{0} \in \mathbb{R}^n$, we have*

$$\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(0)}) = \mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\mathbf{0}) \leq 1. \tag{20}$$

Proof. For $\boldsymbol{\alpha} := \mathbf{0} \in \mathbb{R}^n$, we have $\mathbf{w}(\boldsymbol{\alpha}) = \frac{1}{\lambda n} A\boldsymbol{\alpha} = \mathbf{0} \in \mathbb{R}^d$. Therefore, by definition of the dual objective $\mathcal{D}$ given in (2),

$$0 \leq \mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}) \leq P(\mathbf{w}(\boldsymbol{\alpha})) - \mathcal{D}(\boldsymbol{\alpha}) = 0 - \mathcal{D}(\boldsymbol{\alpha}) \overset{(5),(2)}{\leq} 1. \qquad \square$$

B. Proofs

B.1. Proof of Lemma 2

Indeed, we have

$$\mathcal{D}(\boldsymbol{\alpha} + \gamma \sum_{k=1}^K \Delta\boldsymbol{\alpha}_{[k]}) = \underbrace{-\frac{1}{n}\sum_{i=1}^n \ell_i^*(-\alpha_i - \gamma(\sum_{k=1}^K \Delta\boldsymbol{\alpha}_{[k]})_i)}_{A} - \frac{\lambda}{2} \underbrace{\| \frac{1}{\lambda n} A(\boldsymbol{\alpha} + \gamma \sum_{k=1}^K \Delta\boldsymbol{\alpha}_{[k]}) - \mathbf{b} \|^2}_{B}. \tag{21}$$

Now, let us bound the terms A and B separately. We have

$$\begin{aligned} A &= -\frac{1}{n}\sum_{k=1}^K \left(\sum_{i \in \mathcal{P}_k} \ell_i^*(-\alpha_i - \gamma(\Delta\boldsymbol{\alpha}_{[k]})_i) \right) = -\frac{1}{n}\sum_{k=1}^K \left(\sum_{i \in \mathcal{P}_k} \ell_i^*(-(1-\gamma)\alpha_i - \gamma(\boldsymbol{\alpha} + \Delta\boldsymbol{\alpha}_{[k]})_i) \right) \\ &\geq -\frac{1}{n}\sum_{k=1}^K \left(\sum_{i \in \mathcal{P}_k} (1-\gamma)\ell_i^*(-\alpha_i) + \gamma \ell_i^*(-(\boldsymbol{\alpha} + \Delta\boldsymbol{\alpha}_{[k]})_i) \right). \end{aligned}$$

Where the last inequality is due to Jensen. Now we will bound B.

$$\begin{aligned} B = \| \frac{1}{\lambda n} A(\boldsymbol{\alpha} + \gamma \sum_{k=1}^K \Delta\boldsymbol{\alpha}_{[k]}) - \mathbf{b} \|^2 &= \|\mathbf{w}(\boldsymbol{\alpha}) + \gamma \frac{1}{\lambda n} \sum_{k=1}^K A\Delta\boldsymbol{\alpha}_{[k]}\|^2 = \|\mathbf{w}(\boldsymbol{\alpha})\|^2 + \sum_{k=1}^K 2\gamma \frac{1}{\lambda n} \mathbf{w}(\boldsymbol{\alpha})^T A\Delta\boldsymbol{\alpha}_{[k]} \\ + \gamma(\frac{1}{\lambda n})^2 \gamma \| \sum_{k=1}^K A\Delta\boldsymbol{\alpha}_{[k]}\|^2 &\overset{(9)}{\leq} \|\mathbf{w}(\boldsymbol{\alpha})\|^2 + \sum_{k=1}^K 2\gamma \frac{1}{\lambda n} \mathbf{w}(\boldsymbol{\alpha})^T A\Delta\boldsymbol{\alpha}_{[k]} + \gamma(\frac{1}{\lambda n})^2 \sigma' \sum_{k=1}^K \|A\boldsymbol{\alpha}_{[k]}\|^2. \end{aligned}$$

Plugging A and B into (21) will give us

$$\begin{aligned}
\mathcal{D}(\boldsymbol{\alpha} + \gamma \sum_{k=1}^{K} \Delta\boldsymbol{\alpha}_{[k]}) \geq & -\frac{1}{n}\sum_{k=1}^{K}\left(\sum_{i\in\mathcal{P}_k}(1-\gamma)\ell_i^*(-\alpha_i) + \gamma\ell_i^*(-(\boldsymbol{\alpha}+\Delta\boldsymbol{\alpha}_{[k]})_i)\right) \\
& -\gamma\frac{\lambda}{2}\|\mathbf{w}(\boldsymbol{\alpha})\|^2 - (1-\gamma)\frac{\lambda}{2}\|\mathbf{w}(\boldsymbol{\alpha})\|^2 - \frac{\lambda}{2}\sum_{k=1}^{K}2\gamma\frac{1}{\lambda n}\mathbf{w}(\boldsymbol{\alpha})^T A\Delta\boldsymbol{\alpha}_{[k]} - \frac{\lambda}{2}\gamma(\frac{1}{\lambda n})^2\sigma'\sum_{k=1}^{K}\|A\boldsymbol{\alpha}_{[k]}\|^2 \\
= & \underbrace{-\frac{1}{n}\sum_{k=1}^{K}\left(\sum_{i\in\mathcal{P}_k}(1-\gamma)\ell_i^*(-\alpha_i)\right) - (1-\gamma)\frac{\lambda}{2}\|\mathbf{w}(\boldsymbol{\alpha})\|^2}_{(1-\gamma)\mathcal{D}(\boldsymbol{\alpha})} \\
& + \gamma\sum_{k=1}^{K}\left(-\frac{1}{n}\sum_{i\in\mathcal{P}_k}\ell_i^*(-(\boldsymbol{\alpha}+\Delta\boldsymbol{\alpha}_{[k]})_i) - \frac{1}{K}\frac{\lambda}{2}\|\mathbf{w}(\boldsymbol{\alpha})\|^2 - \frac{1}{n}\mathbf{w}(\boldsymbol{\alpha})^T A\Delta\boldsymbol{\alpha}_{[k]} - \frac{\lambda}{2}\sigma'\|\frac{1}{\lambda n}A\boldsymbol{\alpha}_{[k]}\|^2\right) \\
\overset{(8)}{=} & (1-\gamma)\mathcal{D}(\boldsymbol{\alpha}) + \gamma\sum_{k=1}^{K}\mathcal{G}_k^{\sigma'}(\Delta\boldsymbol{\alpha}_{[k]};\mathbf{w}).
\end{aligned}$$

B.2. Proof of Lemma 1

See (Richtárik & Takáč, 2013).

B.3. Proof of Lemma 3

For sake of notation, we will write $\boldsymbol{\alpha}$ instead of $\boldsymbol{\alpha}^{(t)}$, $\mathbf{w}$ instead of $\mathbf{w}(\boldsymbol{\alpha}^{(t)})$ and $\mathbf{u}$ instead of $\mathbf{u}^{(t)}$.

Now, let us estimate the expected change of the dual objective. Using the definition of the dual update $\boldsymbol{\alpha}^{(t+1)} := \boldsymbol{\alpha}^{(t)} + \gamma\sum_k \Delta\boldsymbol{\alpha}_{[k]}$ resulting in Algorithm 1, we have

$$\begin{aligned}
\mathbb{E}\big[\mathcal{D}(\boldsymbol{\alpha}^{(t)}) - \mathcal{D}(\boldsymbol{\alpha}^{(t+1)})\big] = & \mathbb{E}\Big[\mathcal{D}(\boldsymbol{\alpha}) - \mathcal{D}(\boldsymbol{\alpha} + \gamma\sum_{k=1}^{K}\Delta\boldsymbol{\alpha}_{[k]})\Big] \\
& \text{(by Lemma 2 on the local function } \mathcal{G}_k^{\sigma'}(\boldsymbol{\alpha};\mathbf{w}) \text{ approximating the global objective } \mathcal{D}(\boldsymbol{\alpha})) \\
\leq & \mathbb{E}\Big[\mathcal{D}(\boldsymbol{\alpha}) - (1-\gamma)\mathcal{D}(\boldsymbol{\alpha}) - \gamma\sum_{k=1}^{K}\mathcal{G}_k^{\sigma'}(\Delta\boldsymbol{\alpha}_{[k]}^{(t)};\mathbf{w})\Big] \\
= & \gamma\mathbb{E}\Big[\mathcal{D}(\boldsymbol{\alpha}) - \sum_{k=1}^{K}\mathcal{G}_k^{\sigma'}(\Delta\boldsymbol{\alpha}_{[k]}^{(t)};\mathbf{w})\Big] \\
= & \gamma\mathbb{E}\Big[\mathcal{D}(\boldsymbol{\alpha}) - \sum_{k=1}^{K}\mathcal{G}_k^{\sigma'}(\Delta\boldsymbol{\alpha}_{[k]}^{*};\mathbf{w}) + \sum_{k=1}^{K}\mathcal{G}_k^{\sigma'}(\Delta\boldsymbol{\alpha}_{[k]}^{*};\mathbf{w}) - \sum_{k=1}^{K}\mathcal{G}_k^{\sigma'}(\Delta\boldsymbol{\alpha}_{[k]}^{(t)};\mathbf{w})\Big] \\
& \text{(by the notion of quality (10) of the local solver, as in Assumption 1)} \\
\leq & \gamma\left(\mathcal{D}(\boldsymbol{\alpha}) - \sum_{k=1}^{K}\mathcal{G}_k^{\sigma'}(\Delta\boldsymbol{\alpha}_{[k]}^{*};\mathbf{w}) + \Theta\left(\sum_{k=1}^{K}\mathcal{G}_k^{\sigma'}(\Delta\boldsymbol{\alpha}_{[k]}^{*};\mathbf{w}) - \underbrace{\sum_{k=1}^{K}\mathcal{G}_k^{\sigma'}(\mathbf{0};\mathbf{w})}_{\mathcal{D}(\boldsymbol{\alpha})}\right)\right) \\
= & \gamma(1-\Theta)\left(\underbrace{\mathcal{D}(\boldsymbol{\alpha}) - \sum_{k=1}^{K}\mathcal{G}_k^{\sigma'}(\Delta\boldsymbol{\alpha}_{[k]}^{*};\mathbf{w})}_{C}\right).
\end{aligned} \tag{22}$$

Now, let us upper bound the C term (we will denote by $\Delta\boldsymbol{\alpha}^* = \sum_{k=1}^K \Delta\boldsymbol{\alpha}^*_{[k]}$):

$$\begin{aligned}
C &\overset{(2),(8)}{=} \frac{1}{n}\sum_{i=1}^n \left(\ell_i^*(-\alpha_i - \Delta\boldsymbol{\alpha}_i^*) - \ell_i^*(-\alpha_i)\right) + \frac{1}{n}\mathbf{w}(\boldsymbol{\alpha})^T A\Delta\boldsymbol{\alpha}^* + \sum_{k=1}^K \frac{\lambda}{2}\sigma' \Big\|\frac{1}{\lambda n}A\Delta\boldsymbol{\alpha}^*_{[k]}\Big\|^2 \\
&\le \frac{1}{n}\sum_{i=1}^n \left(\ell_i^*(-\alpha_i - s(u_i-\alpha_i)) - \ell_i^*(-\alpha_i)\right) + \frac{1}{n}\mathbf{w}(\boldsymbol{\alpha})^T As(\mathbf{u}-\boldsymbol{\alpha}) + \sum_{k=1}^K \frac{\lambda}{2}\sigma' \Big\|\frac{1}{\lambda n}As(\mathbf{u}-\boldsymbol{\alpha})_{[k]}\Big\|^2 \\
&\overset{\text{Strong conv.}}{\le} \frac{1}{n}\sum_{i=1}^n \Big(s\ell_i^*(-u_i) + (1-s)\ell_i^*(-\alpha_i) - \frac{\mu}{2}(1-s)s(u_i-\alpha_i)^2 - \ell_i^*(-\alpha_i)\Big) + \frac{1}{n}\mathbf{w}(\boldsymbol{\alpha})^T As(\mathbf{u}-\boldsymbol{\alpha}) \\
&\qquad + \sum_{k=1}^K \frac{\lambda}{2}\sigma' \Big\|\frac{1}{\lambda n}As(\mathbf{u}-\boldsymbol{\alpha})_{[k]}\Big\|^2 \\
&= \frac{1}{n}\sum_{i=1}^n \Big(s\ell_i^*(-u_i) - s\ell_i^*(-\alpha_i) - \frac{\mu}{2}(1-s)s(u_i-\alpha_i)^2\Big) + \frac{1}{n}\mathbf{w}(\boldsymbol{\alpha})^T As(\mathbf{u}-\boldsymbol{\alpha}) + \sum_{k=1}^K \frac{\lambda}{2}\sigma' \Big\|\frac{1}{\lambda n}As(\mathbf{u}-\boldsymbol{\alpha})_{[k]}\Big\|^2.
\end{aligned}$$

The convex conjugate maximal property implies that

$$\ell_i^*(-u_i) = -u_i\mathbf{w}(\boldsymbol{\alpha})^T\mathbf{x}_i - \ell_i(\mathbf{w}(\boldsymbol{\alpha})^T\mathbf{x}_i). \tag{23}$$

Moreover, from the definition of the primal and dual optimization problems (1), (2), we can write the duality gap as

$$G(\boldsymbol{\alpha}) := \mathcal{P}(\mathbf{w}(\boldsymbol{\alpha})) - \mathcal{D}(\boldsymbol{\alpha}) \overset{(1),(2)}{=} \frac{1}{n}\sum_{i=1}^n \left(\ell_i(\mathbf{x}_j^T\mathbf{w}) + \ell_i^*(-\alpha_i) + \mathbf{w}(\boldsymbol{\alpha})^T\mathbf{x}_i\alpha_i\right). \tag{24}$$

Hence,

$$\begin{aligned}
C &\overset{(23)}{\le} \frac{1}{n}\sum_{i=1}^n \left(-su_i\mathbf{w}(\boldsymbol{\alpha})^T\mathbf{x}_i - s\ell_i(\mathbf{w}(\boldsymbol{\alpha})^T\mathbf{x}_i) - s\ell_i^*(-\alpha_i)\underbrace{-s\mathbf{w}(\boldsymbol{\alpha})^T\mathbf{x}_i\alpha_i + s\mathbf{w}(\boldsymbol{\alpha})^T\mathbf{x}_i\alpha_i}_{0} - \frac{\mu}{2}(1-s)s(u_i-\alpha_i)^2\right) \\
&\qquad + \frac{1}{n}\mathbf{w}(\boldsymbol{\alpha})^T As(\mathbf{u}-\boldsymbol{\alpha}) + \sum_{k=1}^K \frac{\lambda}{2}\sigma' \Big\|\frac{1}{\lambda n}As(\mathbf{u}-\boldsymbol{\alpha})_{[k]}\Big\|^2 \\
&= \frac{1}{n}\sum_{i=1}^n \left(-s\ell_i(\mathbf{w}(\boldsymbol{\alpha})^T\mathbf{x}_i) - s\ell_i^*(-\alpha_i) - s\mathbf{w}(\boldsymbol{\alpha})^T\mathbf{x}_i\alpha_i\right) + \frac{1}{n}\sum_{i=1}^n \Big(s\mathbf{w}(\boldsymbol{\alpha})^T\mathbf{x}_i(\alpha_i - u_i) - \frac{\mu}{2}(1-s)s(u_i-\alpha_i)^2\Big) \\
&\qquad + \frac{1}{n}\mathbf{w}(\boldsymbol{\alpha})^T As(\mathbf{u}-\boldsymbol{\alpha}) + \sum_{k=1}^K \frac{\lambda}{2}\sigma' \Big\|\frac{1}{\lambda n}As(\mathbf{u}-\boldsymbol{\alpha})_{[k]}\Big\|^2 \\
&\overset{(24)}{=} -sG(\boldsymbol{\alpha}) - \frac{\mu}{2}(1-s)s\frac{1}{n}\sum_{i=1}^n \|\mathbf{u}-\boldsymbol{\alpha}\|^2 + \frac{\sigma'}{2\lambda}(\frac{s}{n})^2\sum_{k=1}^K \|A(\mathbf{u}-\boldsymbol{\alpha})_{[k]}\|^2.
\end{aligned} \tag{25}$$

Now, the claimed improvement bound (13) follows by plugging (25) into (22).

B.4. Proof of Lemma 4

For general convex functions, the strong convexity parameter is $\mu = 0$, and hence the definition of $R^{(t)}$ becomes

$$R^{(t)} \overset{(14)}{=} \sum_{k=1}^K \|A(\mathbf{u}^{(t)} - \boldsymbol{\alpha}^{(t)})_{[k]}\|^2 \overset{(17)}{\le} \sum_{k=1}^K \sigma_k \|(\mathbf{u}^{(t)} - \boldsymbol{\alpha}^{(t)})_{[k]}\|^2 \overset{\text{Lemma 8}}{\le} \sum_{k=1}^K \sigma_k |\mathcal{P}_k| 4L^2.$$

B.5. Proof of Theorem 6

At first let us estimate expected change of dual feasibility. By using the main Lemma 3, we have

$$\begin{aligned}
\mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(t+1)})] &= \mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(t+1)}) + \mathcal{D}(\boldsymbol{\alpha}^{(t)}) - \mathcal{D}(\boldsymbol{\alpha}^{(t)})] \\
&\overset{(13)}{=} \mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(t)}) - \gamma(1-\Theta)sG(\boldsymbol{\alpha}^{(t)}) + \gamma(1-\Theta)\tfrac{\sigma'}{2\lambda}(\frac{s}{n})^2 R^{(t)} \\
&\overset{(4)}{=} \mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(t)}) - \gamma(1-\Theta)s(\mathcal{P}(\mathbf{w}(\boldsymbol{\alpha}^{(t)})) - \mathcal{D}(\boldsymbol{\alpha}^{(t)})) + \gamma(1-\Theta)\tfrac{\sigma'}{2\lambda}(\frac{s}{n})^2 R^{(t)} \\
&\leq \mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(t)}) - \gamma(1-\Theta)s(\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(t)})) + \gamma(1-\Theta)\tfrac{\sigma'}{2\lambda}(\frac{s}{n})^2 R^{(t)} \\
&\overset{(16)}{\leq} (1-\gamma(1-\Theta)s)\left(\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(t)})\right) + \gamma(1-\Theta)\tfrac{\sigma'}{2\lambda}(\frac{s}{n})^2 4L^2\sigma. \qquad (26)
\end{aligned}$$

Using (26) recursively we have

$$\begin{aligned}
\mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(t)})] &= (1-\gamma(1-\Theta)s)^t\left(\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(0)})\right) + \gamma(1-\Theta)\tfrac{\sigma'}{2\lambda}(\frac{s}{n})^2 4L^2\sigma \sum_{j=0}^{t-1}(1-\gamma(1-\Theta)s)^j \\
&= (1-\gamma(1-\Theta)s)^t\left(\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(0)})\right) + \gamma(1-\Theta)\tfrac{\sigma'}{2\lambda}(\frac{s}{n})^2 4L^2\sigma \frac{1-(1-\gamma(1-\Theta)s)^t}{\gamma(1-\Theta)s} \\
&\leq (1-\gamma(1-\Theta)s)^t\left(\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(0)})\right) + s\frac{4L^2\sigma\sigma'}{2\lambda n^2}. \qquad (27)
\end{aligned}$$

Choice of $s=1$ and $t = t_0 := \max\{0, \lceil \frac{1}{\gamma(1-\Theta)}\log(2\lambda n^2(\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(0)}))/(4L^2\sigma\sigma'))\rceil\}$ will lead to

$$\mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(t)})] \leq (1-\gamma(1-\Theta))^{t_0}\left(\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(0)})\right) + \frac{4L^2\sigma\sigma'}{2\lambda n^2} \leq \frac{4L^2\sigma\sigma'}{2\lambda n^2} + \frac{4L^2\sigma\sigma'}{2\lambda n^2} = \frac{4L^2\sigma\sigma'}{\lambda n^2}. \qquad (28)$$

Now, we are going to show that

$$\forall t \geq t_0 : \mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(t)})] \leq \frac{4L^2\sigma\sigma'}{\lambda n^2(1+\frac{1}{2}\gamma(1-\Theta)(t-t_0))}. \qquad (29)$$

Clearly, (28) implies that (29) holds for $t = t_0$. Now imagine that it holds for any $t \geq t_0$ then we show that it also has to hold for $t+1$. Indeed, using

$$s = \frac{1}{1+\frac{1}{2}\gamma(1-\Theta)(t-t_0)} \in [0,1] \qquad (30)$$

we obtain

$$\begin{aligned}
\mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(t+1)})] &\overset{(26)}{\leq} (1-\gamma(1-\Theta)s)\left(\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(t)})\right) + \gamma(1-\Theta)\tfrac{\sigma'}{2\lambda}(\frac{s}{n})^2 4L^2\sigma \\
&\overset{(29)}{\leq} (1-\gamma(1-\Theta)s)\frac{4L^2\sigma\sigma'}{\lambda n^2(1+\frac{1}{2}\gamma(1-\Theta)(t-t_0))} + \gamma(1-\Theta)\tfrac{\sigma'}{2\lambda}(\frac{s}{n})^2 4L^2\sigma \\
&\overset{(30)}{=} \frac{4L^2\sigma\sigma'}{\lambda n^2}\left(\frac{1+\frac{1}{2}\gamma(1-\Theta)(t-t_0) - \gamma(1-\Theta) + \gamma(1-\Theta)\frac{1}{2}}{(1+\frac{1}{2}\gamma(1-\Theta)(t-t_0))^2}\right) \\
&= \frac{4L^2\sigma\sigma'}{\lambda n^2}\underbrace{\left(\frac{1+\frac{1}{2}\gamma(1-\Theta)(t-t_0) - \frac{1}{2}\gamma(1-\Theta)}{(1+\frac{1}{2}\gamma(1-\Theta)(t-t_0))^2}\right)}_{D}.
\end{aligned}$$

Now, we will upperbound D as follows

$$\begin{aligned}
D &= \frac{1}{1+\frac{1}{2}\gamma(1-\Theta)(t+1-t_0)}\underbrace{\frac{(1+\frac{1}{2}\gamma(1-\Theta)(t+1-t_0))(1+\frac{1}{2}\gamma(1-\Theta)(t-1-t_0))}{(1+\frac{1}{2}\gamma(1-\Theta)(t-t_0))^2}}_{\leq 1} \\
&\leq \frac{1}{1+\frac{1}{2}\gamma(1-\Theta)(t+1-t_0)},
\end{aligned}$$

where in the last inequality we have used the fact that geometric mean is less or equal to arithmetic mean.

If $\overline{\boldsymbol{\alpha}}$ is defined as (19) then we obtain that

$$\begin{aligned}
\mathbb{E}[G(\overline{\boldsymbol{\alpha}})] &= \mathbb{E}\left[G\left(\sum_{t=T_0}^{T-1} \tfrac{1}{T-T_0}\boldsymbol{\alpha}^{(t)}\right)\right] \leq \tfrac{1}{T-T_0}\mathbb{E}\left[\sum_{t=T_0}^{T-1} G\left(\boldsymbol{\alpha}^{(t)}\right)\right] \\
&\overset{(13),(16)}{\leq} \tfrac{1}{T-T_0}\mathbb{E}\left[\sum_{t=T_0}^{T-1}\left(\frac{1}{\gamma(1-\Theta)s}(\mathcal{D}(\boldsymbol{\alpha}^{(t+1)}) - \mathcal{D}(\boldsymbol{\alpha}^{(t)})) + \tfrac{4L^2\sigma\sigma' s}{2\lambda n^2}\right)\right] \\
&= \frac{1}{\gamma(1-\Theta)s}\frac{1}{T-T_0}\mathbb{E}\left[\mathcal{D}(\boldsymbol{\alpha}^{(T)}) - \mathcal{D}(\boldsymbol{\alpha}^{(T_0)})\right] + \tfrac{4L^2\sigma\sigma' s}{2\lambda n^2} \\
&\leq \frac{1}{\gamma(1-\Theta)s}\frac{1}{T-T_0}\mathbb{E}\left[\mathcal{D}(\boldsymbol{\alpha}^{*}) - \mathcal{D}(\boldsymbol{\alpha}^{(T_0)})\right] + \tfrac{4L^2\sigma\sigma' s}{2\lambda n^2}. \qquad (31)
\end{aligned}$$

Now, if $T \geq \lceil \frac{1}{\gamma(1-\Theta)} \rceil + T_0$ such that $T_0 \geq t_0$ we obtain

$$\begin{aligned}
\mathbb{E}[G(\overline{\boldsymbol{\alpha}})] &\overset{(31),(29)}{\leq} \frac{1}{\gamma(1-\Theta)s}\frac{1}{T-T_0}\left(\frac{4L^2\sigma\sigma'}{\lambda n^2(1+\frac{1}{2}\gamma(1-\Theta)(T_0-t_0))}\right) + \frac{4L^2\sigma\sigma' s}{2\lambda n^2} \\
&= \frac{4L^2\sigma\sigma'}{\lambda n^2}\left(\frac{1}{\gamma(1-\Theta)s}\frac{1}{T-T_0}\frac{1}{1+\frac{1}{2}\gamma(1-\Theta)(T_0-t_0)} + \frac{s}{2}\right). \qquad (32)
\end{aligned}$$

Choosing

$$s = \frac{1}{(T-T_0)\gamma(1-\Theta)} \in [0,1] \qquad (33)$$

gives us

$$\mathbb{E}[G(\overline{\boldsymbol{\alpha}})] \overset{(32),(33)}{\leq} \frac{4L^2\sigma\sigma'}{\lambda n^2}\left(\frac{1}{1+\frac{1}{2}\gamma(1-\Theta)(T_0-t_0)} + \frac{1}{(T-T_0)\gamma(1-\Theta)}\frac{1}{2}\right). \qquad (34)$$

To have right hand side of (34) smaller then ϵ_G it is sufficient to choose T_0 and T such that

$$\begin{aligned}
\frac{4L^2\sigma\sigma'}{\lambda n^2}\left(\frac{1}{1+\frac{1}{2}\gamma(1-\Theta)(T_0-t_0)}\right) &\leq \frac{1}{2}\epsilon_G, \qquad (35) \\
\frac{4L^2\sigma\sigma'}{\lambda n^2}\left(\frac{1}{(T-T_0)\gamma(1-\Theta)}\frac{1}{2}\right) &\leq \frac{1}{2}\epsilon_G. \qquad (36)
\end{aligned}$$

Hence of if

$$\begin{aligned}
t_0 + \frac{2}{\gamma(1-\Theta)}\left(\frac{8L^2\sigma\sigma'}{\lambda n^2\epsilon_G} - 1\right) &\leq T_0, \\
T_0 + \frac{4L^2\sigma\sigma'}{\lambda n^2\epsilon_G\gamma(1-\Theta)} &\leq T,
\end{aligned}$$

then (35) and (36) are satisfied.