Nesterov Meets Robust Multitask Learning Twice

Yifan Kang YIFANK@CLEMSON.EDU
Kai Liu KAIL@CLEMSON.EDU
Clemson University

Abstract

In this paper, we study temporal multitask learning problem where we impose smoothness constraint on time-series weights. Besides, to select important features, group lasso is introduced. Moreover, the regression loss in each time frame is non-squared to alleviate the influence of various scales of noise in each task, in addition to the nuclear norm for low-rank property. We first formulate the objective as a max-min problem, where the dual variable can be optimized via accelerated dual ascent method, while the primal variable can be solved via *smoothed Fast Iterative Shrinkage-Thresholding Algorithm* (S-FISTA). We provide convergence analysis of the proposed method and experiments demonstrate its effectiveness.

1. Multi-task Learning with Multiple Regularizers

Multi-task learning aims to improve the generalization performance by learning multiple related tasks together and exploring the shared features among tasks. It has received a lot of interests and has been successfully applied to lots of applications including gene data analysis [11], breast cancer classification [22], and disease progression prediction [24, 25]. Most existing multi-task feature learning models can be formulated as a regularized optimization problem and they usually focus on how to design a good regularizer to capture the underlying shared features among tasks; examples include group lasso multi-task feature learning [7–9, 15, 21], low rank constraint (including its convex envelope nuclear minimization) [4, 12, 13, 16, 18], etc. In cases where spatio-temporal structure is considered, weights in close region or time period should be similar [6, 23, 24]. Besides, most existing multi-task feature learning models simply assume a common noise level for all tasks, which may not hold in real applications [10, 14]. Moreover, theoretical analysis [15] shows that, to achieve the optimal parameter estimation error bounds, the regularized parameter should be chosen in proportion to the maximum standard deviations of the noise for all tasks. In practice, the standard deviations of the noise are unknown or very difficult to estimate, which makes the parameter tuning quite challenging [10]. In this paper, we impose various constraints on multitask learning to solve:

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times m}} \mathcal{J}(\mathbf{W}) = \underbrace{\sum_{i=1}^{m} \|\mathbf{X}_i\boldsymbol{w}_i - \boldsymbol{y}_i\|_2 + \lambda\|\mathbf{W}\|_{2,1}}_{h(\mathbf{W})} + \underbrace{\frac{\gamma}{2}\|\mathbf{W}\|_F^2 + \frac{\zeta}{2}\sum_{i=1}^{m-1}\|\boldsymbol{w}_i - \boldsymbol{w}_{i+1}\|_2^2}_{f(\mathbf{W})} + \underbrace{\rho\|\mathbf{W}\|_*}_{g(\mathbf{W})} \tag{1}$$

where $\mathbf{X}_i \in \mathbb{R}^{n_i\times d}, \boldsymbol{y}_i \in \mathbb{R}^{n_i}$, $\|\mathbf{W}\|_{2,1} = \sum_{i=1}\|\mathbf{W}(i,:)\|_2$, $\|\mathbf{X}\|_* = \sum_{i=1}\sigma_i(\mathbf{W})$. $\|\mathbf{X}_i\boldsymbol{w}_i - \boldsymbol{y}_i\|_2$ measures the robust loss in each time stamp due to the potential different noise levels of all tasks for calibration (as we will discuss later, other *non-squared* norm, such as ℓ_1-norm also applies). $\|\mathbf{W}\|_{2,1}$ is to select important features, enforcing row-wise sparsity, i.e., it encourages all-zero-value rows in

$\mathbf{W}$. $\|\boldsymbol{w}_i - \boldsymbol{w}_{i+1}\|_2^2$ is to introduce smoothness among each temporal task. A combination of $\|\mathbf{W}\|_*$ and $\|\mathbf{W}\|_F^2$ follows the idea of *elastic net* [26] in least squares, which plays a role in imposing low-rank constraint as $\|\mathbf{W}\|_F^2 = \sum_{i=1} \sigma_i^2(\mathbf{W})$. The main challenge to optimize Eq. (1) is it contains multiple non-smooth terms in $h(\mathbf{W})$ and $g(\mathbf{W})$.

2. Smoothing

Sub-gradient descent method is applicable for Eq. (1) but known to suffer from slow convergence which is non-monotonic. Therefore in this paper we turn to seek for methods with faster convergence rate and guarantee. Inspired by [17], we can smooth both $\|\mathbf{X}_i\boldsymbol{w}_i - \boldsymbol{y}_i\|_2$ and $\|\mathbf{W}\|_{2,1}$ terms.

For sake of further analysis, we begin with the following definition:

Definition 1 (smoothable function) *[1] A convex function h is called (α, β)-**smoothable** $(\alpha, \beta > 0)$ if for any $\mu > 0$ there exists a convex differentiable function h_μ such that the following holds:*

- *$h_\mu(\boldsymbol{x}) \leq h(\boldsymbol{x}) \leq h_\mu(\boldsymbol{x}) + \beta\mu$ for any $\boldsymbol{x}$.*
- *$h_\mu(\boldsymbol{x})$ is $\frac{\alpha}{\mu}$-smooth.*

The function h_μ is called a $\frac{1}{\mu}$-smooth approximation of h with parameters (α, β).

Theorem 2 *For $h(\boldsymbol{x}) = \|\boldsymbol{x}\|_2$, function*

$$h_\mu(\boldsymbol{x}) = \begin{cases} \frac{1}{2\mu}\|\boldsymbol{x}\|_2^2, & \|\boldsymbol{x}\|_2 \leq \mu, \\ \|\boldsymbol{x}\|_2 - \frac{\mu}{2}, & \|\boldsymbol{x}\|_2 > \mu, \end{cases} \tag{2}$$

is a $\frac{1}{\mu}$-smooth approximation of h with parameters $(1, 0.5)$ [1] [1].

Based on Theorem 2, we can smooth $\|\mathbf{W}\|_{2,1}$ by each row and the smoothed term $\|\mathbf{W}\|_{2,1}^\mu$ is $(1, \frac{d}{2})$-smoothable. Moreover, the following theorem provides an exact approximation for $\|\mathbf{X}_i\boldsymbol{w}_i - \boldsymbol{y}_i\|_2$:

Theorem 3 *Let $\mathcal{A}$ be a linear transformation and h be a convex function and define:*

$$q(\boldsymbol{x}) = h(\mathcal{A}(\boldsymbol{x}) + b), \tag{3}$$

Assume h_μ is a $\frac{1}{\mu}$-smooth approximation of h with parameters (α, β), then $q_\mu(x) = h_\mu(\mathcal{A}(\boldsymbol{x}) + b)$ is a $\frac{1}{\mu}$-smooth approximation of q with parameters $(\alpha\|\mathcal{A}\|_2^2, \beta)$.

Thus, by smoothing $\|\mathbf{X}_i\boldsymbol{w}_i - \boldsymbol{y}_i\|_2$ and Theorem 2, $\|\mathbf{X}_i\boldsymbol{w}_i - \boldsymbol{y}_i\|_2^\mu$ is $(\|\mathbf{X}_i\|_2^2, \frac{1}{2})$-smoothable.

Based on the analysis aforementioned, $h_\mu := \sum_{i=1}^m \|\mathbf{X}_i\boldsymbol{w}_i - \boldsymbol{y}_i\|_2^\mu + \lambda\|\mathbf{W}\|_{2,1}^\mu$ is a $\frac{1}{\mu}$-smooth approximation of $h = \sum_{i=1}^m \|\mathbf{X}_i\boldsymbol{w}_i - \boldsymbol{y}_i\|_2 + \lambda\|\mathbf{W}\|_{2,1}$ with parameters $(\lambda + \max_i \|\mathbf{X}_i\|_2^2, \frac{m+\lambda d}{2})$.

1. One can verify that this can be derived from *Moreau Envelope* which can get the tightest bound on β. Other options say $h_\mu(\boldsymbol{x}) = \sqrt{\|\boldsymbol{x}\|_2^2 + \mu^2} - \mu$ also works with parameters $(1, 1)$.

We can write the smoothing function of $\mathcal{J}$ in Eq. (1) as:

$$\mathcal{J}_\mu(\mathbf{W}) = \underbrace{h_\mu(\mathbf{W}) + f(\mathbf{W})}_{F_\mu(\mathbf{W})} + g(\mathbf{W}), \tag{4}$$

where F_μ is a smooth function and g is non-smooth but convex. Therefore, we can follow the well-known FISTA [1] to update 4. Obviously, from $h(\mathbf{W}) - \frac{(m+\lambda d)\mu}{2} \leq h_\mu(\mathbf{W}) \leq h(\mathbf{W})$ we have $h(\mathbf{W}) - h^* \leq h_\mu(\mathbf{W}) - h_\mu^* + \frac{(m+\lambda d)\mu}{2}$. To achieve $h(\mathbf{W}) - h^* \leq \delta$, we need $h_\mu(\mathbf{W}) - h_\mu^* \leq \delta - \frac{(m+\lambda d)\mu}{2}$, thus μ should be no more than $\frac{2\delta}{m+\lambda d}$. In case we require high precision solution where δ is small, μ therefore should be set sufficiently small, but as later analysis in Theorem 5 indicates, it would be very time consuming (L is inversely proportional to μ).

Therefore, we propose a new method for acceleration from the perspective of duality.

3. Dual Accelerated Method

Let $\boldsymbol{z}_i = \mathbf{X}_i \boldsymbol{w}_i - \boldsymbol{y}_i$, Eq. (1) can be formulated as a Lagrange function as:

$$\mathcal{L}(\mathbf{W}, \boldsymbol{z}, \theta) = \sum_{i=1}^{m} \|\boldsymbol{z}_i\|_2 + \lambda \|\mathbf{W}\|_{2,1} + f(\mathbf{W}) + g(\mathbf{W}) + \sum_{i=1}^{m} \langle \theta_i, \mathbf{X}_i \boldsymbol{w}_i - \boldsymbol{y}_i - \boldsymbol{z}_i \rangle, \tag{5}$$

where $\theta_i \in \mathbb{R}^{n_i}$. By minimizing $\mathcal{L}(\mathbf{W}, \boldsymbol{z}, \theta)$ w.r.t $\mathbf{W}$ and $\boldsymbol{z}$, we obtain the dual problem by:

$$\tilde{\mathcal{D}}(\theta) = \min_{\mathbf{W}} \left\{ \lambda \|\mathbf{W}\|_{2,1} + f(\mathbf{W}) + g(\mathbf{W}) + \sum_{i=1}^{m} \langle \theta_i, \mathbf{X}_i \boldsymbol{w}_i - \boldsymbol{y}_i \rangle \right\} + \sum_{i=1}^{m} \min_{\boldsymbol{z}_i} \left\{ \|\boldsymbol{z}_i\|_2 - \langle \theta_i, \boldsymbol{z}_i \rangle \right\}. \tag{6}$$

By making use of the fact that

$$\min_{\boldsymbol{z}_i} \left\{ \|\boldsymbol{z}_i\|_2 - \langle \theta_i, \boldsymbol{z}_i \rangle \right\} = \begin{cases} 0, & \|\theta_i\|_2 \leq 1, \\ -\infty, & \text{otherwise}, \end{cases} \tag{7}$$

we obtain the dual problem as: $\max_\theta \mathcal{D}(\theta), \text{ s.t. } \|\theta_i\|_2 \leq 1$ where

$$\mathcal{D}(\theta) = \min_{\mathbf{W}} \mathcal{J}(\mathbf{W}; \theta) = \min_{\mathbf{W}} \left\{ \lambda \|\mathbf{W}\|_{2,1} + f(\mathbf{W}) + g(\mathbf{W}) + \sum_{i=1}^{m} \langle \theta_i, \mathbf{X}_i \boldsymbol{w}_i - \boldsymbol{y}_i \rangle \right\}. \tag{8}$$

It is worth noting that if we change the l_2 norm to general l_p norm in the $\|\mathbf{X}_i \boldsymbol{w}_i - \boldsymbol{y}_i\|_2$ term in Eq. (1), then with minor revision, Eq. (7) still holds:

$$\min_{\boldsymbol{z}_i} \left\{ \|\boldsymbol{z}_i\|_p - \langle \theta_i, \boldsymbol{z}_i \rangle \right\} = \begin{cases} 0, & \|\theta_i\|_q \leq 1, \\ -\infty, & \text{otherwise}, \end{cases} \tag{9}$$

where $\|\cdot\|_q$ is the dual norm of $\|\cdot\|_p$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$. The generalized dual problem becomes $\max_\theta \mathcal{D}(\theta), \text{ s.t. } \|\theta_i\|_q \leq 1$, where $\mathcal{D}(\theta)$ is still the same defined in Eq. (8).

Theorem 4 *The optimization problem in Eq. (8) has a unique solution* $\mathbf{W}(\theta)$ *due to it is strongly convex w.r.t* $\mathbf{W}$. *Moreover,* $\mathcal{D}(\theta)$ *is continuously differentiable and L-Lipschitz continuous gradient. Specifically, the gradient of* $\mathcal{D}(\theta)$ *is*

$$\nabla\mathcal{D}(\theta) = [\mathbf{X}_1\boldsymbol{w}_1(\theta) - \boldsymbol{y}_1; \mathbf{X}_2\boldsymbol{w}_2(\theta) - \boldsymbol{y}_2; \ldots; \mathbf{X}_m\boldsymbol{w}_m(\theta) - \boldsymbol{y}_m], \tag{10}$$

where $\boldsymbol{w}_i(\theta)$ *is the* i*-th column of* $\mathbf{W}(\theta)$ *and Lipschitz continuous gradient is given by*

$$L_{\mathcal{D}} = \frac{\max_i \|\mathbf{X}_i\|_2^2}{\gamma}. \tag{11}$$

Due to the benign property of objective function in Eq. (1), strong duality holds and to find optimal $\mathbf{W}$, one can turn to optimize:

$$\max_{\theta} \mathcal{D}(\theta), \quad s.t. \ \ \|\theta_i\|_2 \leq 1, \ i \in [m], \tag{12}$$

by utilizing FISTA (either constant stepsize $\frac{1}{L_{\mathcal{D}}}$ or backtracking). Notice that the dual problem for $\max_{\theta} \mathcal{D}(\theta), s.t. \ \ \|\theta_i\|_2 \leq 1$ is a maximization problem. Therefore, the gradient projection step and the line search criterion are modified accordingly. We present the pseudo codes in Algorithm 1:

Algorithm 1: Accelerated Dual Ascent with Backtracking

Data: $\mathbf{X}_i, \boldsymbol{y}_i, \lambda, \gamma, \epsilon, \rho$
Result: Optimal $\theta, \mathbf{W}$ to Eq. (5)
initialization $\theta^1 = \theta^0, \eta_0 > 0, \tau > 1, t_1 = t_0 = 1$;
for $k = 1, 2, \ldots$ **do**
  $\alpha_k = \frac{t_{k-1}-1}{t_k}$;
  $\nu^k = \theta^k + \alpha_k(\theta^k - \theta^{k-1})$;
  while *true* **do**
    $\tilde{\nu}^k = [\tilde{\nu}_1^k; \tilde{\nu}_2^k; \ldots; \tilde{\nu}_m^k] = \nu^k + \frac{1}{\eta_k}\nabla\mathcal{D}(\nu^k)$;
    $\theta^k = [\frac{\tilde{\nu}_1^k}{\max(1, \|\tilde{\nu}_1^k\|)}; \frac{\tilde{\nu}_2^k}{\max(1, \|\tilde{\nu}_2^k\|)}; \ldots; \frac{\tilde{\nu}_m^k}{\max(1, \|\tilde{\nu}_m^k\|)}]$;
    % Project into feasible domain $\|\theta_i\| \leq 1$
    if $\mathcal{D}(\theta^k) \geq \mathcal{D}(\nu^k) + \langle\nabla\mathcal{D}(\nu^k), \theta^k - \nu^k\rangle - \frac{\eta_k}{2}\|\theta^k - \nu^k\|^2$ **then**
      break;
    else
      $\eta_k = \tau \cdot \eta_k$;
    end
  end
  Solve $\mathbf{W}(\theta^k)$ in Eq. (8) via Algorithm 2;
  $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
end

Here we discuss the step of projecting θ^k into the feasible domain $\|\theta_i\| \leq 1$. If we are to use general l_p norm instead of l_2 norm for the first term in Eq. (1), θ^k should be projected into the unit ball endowed with the dual (l_q) norm . For a couple of specific p and q settings, there exists simple

solutions to the projection. For example, the $p = q = 2$ case in Algorithm 1. Another often seen case is when $p = 1$, while $q = \infty$, it is straightforward to verify:

$$\theta_i^k(j) = \frac{\tilde{\nu}_i^k(j)}{\max(1, |\tilde{\nu}_i^k(j)|)},$$

where $\theta_i^k(j), \tilde{\nu}_i^k(j)$ denote the j-th entry of the i-th section of $\theta, \tilde{\nu}$ in k-th iteration. Though a closed-form solution is not available for $p = \infty, q = 1$, [5] gives a fast algorithm with $O(N)$ *observed complexity*, while several other algorithms have similar result as well.
For other cases, the projection problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2, \ \text{ s.t. } \ \|\boldsymbol{x}\|_q \leq 1 \tag{13}$$

can be reformulated into

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f_0(\boldsymbol{x}), \ \text{ s.t. } \ \frac{1}{q}(\|\boldsymbol{x}\|_q^q - 1) \leq 0, \tag{14}$$

where $f_0(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$. The Lagrange dual of Eq. (14) is

$$\max_{\mu\geq 0} \ g(\mu) \triangleq \inf_{\boldsymbol{x}\in\mathbb{R}^d} \mathcal{L}_q(\boldsymbol{x}, \mu), \quad \text{where} \quad \mathcal{L}_q(\boldsymbol{x}, \mu) = f_0(\boldsymbol{x}) + \frac{\mu}{q}\left(\sum_{i=1}^d |x_i|^q - 1\right). \tag{15}$$

If $q > 1$, Eq. (15) is a convex optimization problem and readily solvable by using algorithms like Newton's method. In the non-convex cases where $0 < q < 1$, it can be solved with bisection methods [3, 19] or iterative re-weighted l_1-ball projection (IRBP) algorithms [19, 20].

Theorem 5 (Theorem 4.4 [2]) *Let $\{\theta^k\}$ be the sequence generated by Algorithm 1 and θ^* be an optimal solution for the problem in Eq. (8). Then for all $k \geq 1$, we have*

$$\mathcal{D}(\theta^*) - \mathcal{D}(\theta^k) \leq \frac{2\tau L_{\mathcal{D}}\|\theta^* - \theta^0\|^2}{(k+1)^2}, \tag{16}$$

where θ^0 and τ is defined in Algorithm 1, $L_{\mathcal{D}}$ is determined in Eq. (11).

Algorithm 2: Accelerated Gradient Descent Algorithm for Eq. (8)

Data: $\mathbf{X}_i, \boldsymbol{y}_i, \lambda, \gamma, \epsilon, \rho, \mu, \theta$
Result: Optimal $\mathbf{W}$ to Eq. (8)
initialization $\mathbf{V}_1 = \mathbf{W}_0, L = 4\zeta + \gamma + \frac{\lambda}{\mu}, t_1 = 1$;
for $k = 1, 2, \ldots$ **do**
  $\mathbf{W}_k = SVT(\mathbf{V}_k - \frac{1}{L}\nabla\mathcal{J}_\mu(\mathbf{V}_k; \theta), \frac{\rho}{L})$; % SVT is for 'Singular Value Thresholding'
  $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
  $\alpha_k = \frac{t_k - 1}{t_{k+1}}$;
  $\mathbf{V}_{k+1} = \mathbf{W}^k + \alpha_k(\mathbf{W}_k - \mathbf{W}_{k-1})$;
end

It boils down to solve Eq. (8) given certain θ. As we discussed earlier, it contains two non-smooth tern $\ell_{2,1}$-norm and nuclear norm. For sake of simplicity, we smooth $\ell_{2,1}$-norm. Apparently $\mathcal{J}_\mu(\mathbf{W};\theta) := \lambda\|\mathbf{W}\|_{2,1}^\mu + f(\mathbf{W}) + g(\mathbf{W}) + \sum_{i=1}^m \langle \theta_i, \mathbf{X}_i\boldsymbol{w}_i - \boldsymbol{y}_i\rangle$ is $L := 4\zeta + \gamma + \frac{\lambda}{\mu}$ strongly smooth function, where $\|\mathbf{W}\|_{2,1}^\mu$ denotes the smoothing function of $\|\mathbf{W}\|_{2,1}$ for each row in $\mathbf{W}$. The above analysis suggests we can make use of constant stepsize ($\frac{1}{L}$) FISTA to obtain the solution which is almost of the same time consumption compared with backtracking line search.

Theorem 6 *Let $\{\mathbf{W}^k\}$ be the sequences generated by Algorithm 1, and let $\mathbf{W}^*$ be any optimal solutions for the problems in Eq. (1). Then for all $k \geq 1$, we have:*

$$\|\mathbf{W}^k - \mathbf{W}^*\|_F \leq \sqrt{\frac{4\tau L_{\mathcal{D}}\|\theta^* - \theta^0\|^2}{\gamma(k+1)^2} + \frac{d\lambda\mu}{\gamma}}. \tag{17}$$

The dual method, though it also entails smoothing to solve Eq. (8) where $\mu \leq \frac{2\delta}{\lambda d}$, still it is significantly larger than pure smoothing method where $\mu \leq \frac{2\delta}{\lambda d+m}$. Thus the time consumption burden can be alleviated where high precision δ is required. For multitask problem where the number of tasks (m) is very large, dual accelerated method will be superior to the counterpart.

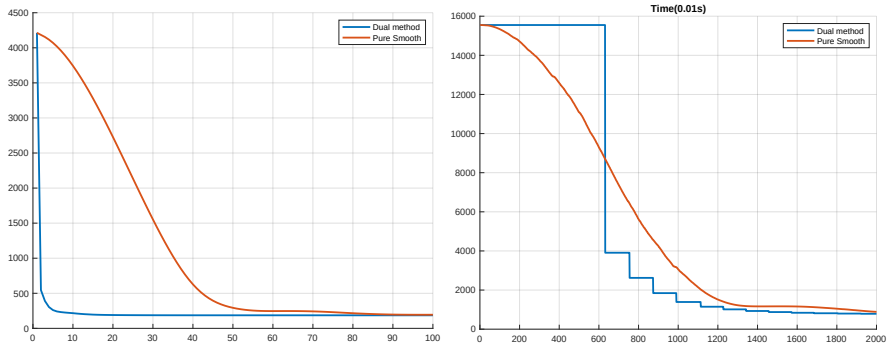

Figure 1: We compare dual vs. pure smoothing methods in terms of objective change with iteration (left) and time consumption (right) on two various settings. Duality method needs very few iterations to converge, and takes less time for high precision solution.

Theorem 7 *Assume $\mathcal{J}(\mathbf{W};\theta)$ has bounded level sets, let $\epsilon \in (0, \bar{\epsilon})$ for some fixed $\bar{\epsilon} > 0$ and $\{\mathbf{W}^k(\theta)\}$ be the sequences generated by Algorithm 2 given certain θ with smoothing parameter*

$$\mu = \sqrt{\frac{2}{d}}\frac{\epsilon}{\sqrt{\alpha\beta} + \sqrt{\alpha\beta + L_f\epsilon}},$$

where $L_f = \gamma + 4\zeta, \alpha = \lambda, \beta = \frac{d\lambda}{2}$. Then for any k satisfying

$$k \geq \frac{2\lambda\sqrt{d\Gamma} + \sqrt{2L_f\Gamma\epsilon}}{\epsilon},$$

where $\Gamma = (\|\mathbf{W}^0(\theta)\|_F + R_{\mathcal{J}(\mathbf{W}^0;\theta)+\frac{\bar{\epsilon}}{2}})^2$, it holds that $\mathcal{J}(\mathbf{W}^k;\theta) - \mathcal{J}(\mathbf{W}^*;\theta) \leq \epsilon$.

4. Conclusion

We study the robust multitask learning by adding various constraints to impose low-rank property and smoothness. We derive a dual optimization problem with a piecewise sphere constraint, which enables us to develop fast dual optimization algorithms. We also provide a detailed convergence analysis for the proposed dual optimization algorithm. Empirical studies demonstrate the dual method quickly converges and it is more efficient than the primal optimization algorithm.

References

[1] Amir Beck. *First-order methods in optimization*. SIAM, 2017.

[2] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[3] Lin Chen, Xue Jiang, Xingzhao Liu, Thiagalingam Kirubarajan, and Zhixin Zhou. Outlier-robust moving object and background decomposition via structured ℓ_p-regularized low-rank representation. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(4): 620–638, 2021. doi: 10.1109/TETCI.2019.2935747.

[4] Bin Cheng, Guangcan Liu, Jingdong Wang, Zhongyang Huang, and Shuicheng Yan. Multi-task low-rank affinity pursuit for image segmentation. In *2011 International conference on computer vision*, pages 2439–2446. IEEE, 2011.

[5] Laurent Condat. Fast projection onto the simplex and the l_1 ball. *Mathematical Programming*, 158(1):575–585, 2016.

[6] Lan Du, Penghui Wang, Hongwei Liu, Mian Pan, Feng Chen, and Zheng Bao. Bayesian spatiotemporal multitask learning for radar hrrp target recognition. *IEEE Transactions on Signal Processing*, 59(7):3182–3196, 2011.

[7] Pinghua Gong, Jieping Ye, and Chang-shui Zhang. Multi-stage multi-task feature learning. *Advances in neural information processing systems*, 25, 2012.

[8] Pinghua Gong, Jieping Ye, and Changshui Zhang. Robust multi-task feature learning. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 895–903, 2012.

[9] Pinghua Gong, Jieping Ye, and Changshui Zhang. Multi-stage multi-task feature learning. *The Journal of Machine Learning Research*, 14(1):2979–3010, 2013.

[10] Pinghua Gong, Jiayu Zhou, Wei Fan, and Jieping Ye. Efficient multi-task feature learning with calibration. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 761–770, 2014.

[11] Seyoung Kim and Eric P Xing. Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eqtl mapping. *Annals of Applied Statistics*, 2012.

[12] Abhishek Kumar and Hal Daume III. Learning task grouping and overlap in multi-task learning. *arXiv preprint arXiv:1206.6417*, 2012.

[13] Congyan Lang, Guangcan Liu, Jian Yu, and Shuicheng Yan. Saliency detection by multitask sparsity pursuit. *IEEE transactions on image processing*, 21(3):1327–1338, 2011.

[14] Han Liu, Lie Wang, and Tuo Zhao. Multivariate regression with calibration. *Advances in neural information processing systems*, 27, 2014.

[15] Karim Lounici, Massimiliano Pontil, Alexandre B Tsybakov, and Sara Van De Geer. Taking advantage of sparsity in multi-task learning. *arXiv preprint arXiv:0903.1468*, 2009.

[16] David Mateos-Núñez, Jorge Cortés, and Jorge Cortes. Distributed optimization for multi-task learning via nuclear-norm approximation. *IFAC-PapersOnLine*, 48(22):64–69, 2015.

[17] Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103:127–152, 2005.

[18] Ting Kei Pong, Paul Tseng, Shuiwang Ji, and Jieping Ye. Trace norm regularization: Reformulations, algorithms, and multi-task learning. *SIAM Journal on Optimization*, 20(6):3465–3489, 2010.

[19] Joong-Ho Won, Kenneth Lange, and Jason Xu. A unified analysis of convex and non-convex l_p-ball projection problems. *Optimization Letters*, 17(5):1133–1159, 2023.

[20] Xiangyu Yang, Jiashan Wang, and Hao Wang. Towards an efficient approach for the nonconvex ℓ_p ball projection: Algorithm and analysis, 2022.

[21] Xiaolin Yang, Seyoung Kim, and Eric Xing. Heterogeneous multitask learning with joint sparsity constraints. *Advances in neural information processing systems*, 22, 2009.

[22] Yu Zhang, Dit-Yan Yeung, and Qian Xu. Probabilistic multi-task feature selection. *Advances in neural information processing systems*, 23, 2010.

[23] Liang Zhao, Qian Sun, Jieping Ye, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. Multi-task learning for spatio-temporal event forecasting. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1503–1512, 2015.

[24] Jiayu Zhou, Lei Yuan, Jun Liu, and Jieping Ye. A multi-task learning formulation for predicting disease progression. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 814–822, 2011.

[25] Jiayu Zhou, Jun Liu, Vaibhav A Narayan, and Jieping Ye. Modeling disease progression via fused sparse group lasso. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1095–1103, 2012.

[26] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

Appendix A. Proof of Theorem 3

Since h_μ is $\frac{\alpha}{\mu}$-smooth, we have

$$\|\nabla h_\mu(\boldsymbol{x}) - \nabla h_\mu(\boldsymbol{y})\|_2 \leq \frac{\alpha}{\mu}\|\boldsymbol{x} - \boldsymbol{y}\|_2. \tag{18}$$

Then

$$\begin{aligned}\|\nabla q_\mu(\boldsymbol{x}) - \nabla q_\mu(\boldsymbol{y})\|_2 =& \|\nabla h_\mu(\mathcal{A}^T\mathcal{A}(\boldsymbol{x}) + b) - \mathcal{A}^T\nabla h_\mu(\mathcal{A}(\boldsymbol{y}) + b)\|_2 \\ \leq& \|\mathcal{A}^T\|_2\|\nabla h_\mu(\mathcal{A}(\boldsymbol{x}) + b) - \nabla h_\mu(\mathcal{A}(\boldsymbol{y}) + b)\|_2 \\ \leq& \|\mathcal{A}\|_2 \cdot \frac{\alpha}{\mu}\|(\mathcal{A}(\boldsymbol{x}) + b) - (\mathcal{A}(\boldsymbol{y}) + b)\|_2 \\ \leq& \frac{\alpha\|\mathcal{A}\|_2^2}{\mu}\|\boldsymbol{x} - \boldsymbol{y}\|_2.\end{aligned} \tag{19}$$

Note that $q_\mu(\boldsymbol{x}) = h_\mu(\mathcal{A}(\boldsymbol{x}) + b) \leq h(\mathcal{A}(\boldsymbol{x}) + b) = q(\boldsymbol{x}) \leq h_\mu(\mathcal{A}(\boldsymbol{x}) + b) + \beta\mu = q_\mu(\boldsymbol{x}) + \beta\mu$, so q_μ is a $\frac{1}{\mu}$-smooth approximation of q with parameters $(\alpha\|\mathcal{A}\|_2^2, \beta)$.

Appendix B. Proof of Theorem 4

Using the following notations:

$$\begin{aligned}\theta &= [\theta_1; \cdots; \theta_m], \\ \boldsymbol{y} &= [\boldsymbol{y}_1; \cdots; \boldsymbol{y}_m], \\ U(\theta) &= [\mathbf{X}_1^T\theta_1, \cdots, \mathbf{X}_m^T\theta_m], \\ \mathcal{F}(\mathbf{W}) &= \lambda\|\mathbf{W}\|_{2,1} + f(\mathbf{W}) + g(\mathbf{W}),\end{aligned} \tag{20}$$

we can rewrite $\mathcal{J}$ in Eq. (8) as

$$\mathcal{J}(\mathbf{W}; \theta) = \mathcal{F}(\mathbf{W}) + \langle \mathbf{W}, U(\theta)\rangle - \theta^T\boldsymbol{y}, \tag{21}$$

thus for Eq. (8), we have

$$\begin{aligned}\mathcal{D}(\theta) =& \min_{\mathbf{W}} \mathcal{J}(\mathbf{W}; \theta) = \min_{\mathbf{W}}\{\mathcal{F}(\mathbf{W}) + \langle \mathbf{W}, U(\theta)\rangle - \theta^T\boldsymbol{y}\} \\ =& -\max_{\mathbf{W}}\{-\langle \mathbf{W}, U(\theta)\rangle - \mathcal{F}(\mathbf{W})\} - \theta^T\boldsymbol{y} \\ =& -\mathcal{F}^*(-U(\theta))) - \theta^T\boldsymbol{y},\end{aligned} \tag{22}$$

where $\mathcal{F}^*$ is the conjugate function of $\mathcal{F}$. Since $\mathcal{F}$ is strongly convex with parameter γ, $\mathcal{F}^*$ is strongly smooth and thus has an L-Lipschitz continuous gradient with parameter $\frac{1}{\gamma}$.

Assume $\mathbf{W}(\theta)$ is the optimal solution for the above equation, then:

$$\mathbf{W}(\theta) = \arg\max_{\mathbf{W}} -\langle \mathbf{W}, U(\theta)\rangle - \mathcal{F}(\mathbf{W}), \tag{23}$$

therefore $-U(\theta) \in \partial\mathcal{F}(\mathbf{W}(\theta))$. Due to the fact that $\mathcal{F}$ is convex, we know:

$$\mathbf{W}(\theta) = \nabla\mathcal{F}^*(-U(\theta)). \tag{24}$$

Recall $\mathcal{D}(\theta) = \min_{\mathbf{W}} \left\{ \lambda \|\mathbf{W}\|_{2,1} + f(\mathbf{W}) + g(\mathbf{W}) + \sum_{i=1}^{m} \langle \theta_i, \mathbf{X}_i \boldsymbol{w}_i - \boldsymbol{y}_i \rangle \right\}$, we have:

$$\nabla \mathcal{D}(\theta) = [\mathbf{X}_1 \boldsymbol{w}_1(\theta) - \boldsymbol{y}_1; \mathbf{X}_2 \boldsymbol{w}_2(\theta) - \boldsymbol{y}_2; \ldots; \mathbf{X}_m \boldsymbol{w}_m(\theta) - \boldsymbol{y}_m] = \mathbf{X}\boldsymbol{s}(\theta) - \boldsymbol{y}, \tag{25}$$

where $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & & \\ & \ddots & \\ & & \mathbf{X}_m \end{bmatrix}$ and $\boldsymbol{s}(\theta) = \begin{bmatrix} \nabla \mathcal{F}_1^*(-U(\theta)) \\ \vdots \\ \nabla \mathcal{F}_m^*(-U(\theta)) \end{bmatrix}$.

As discussed earlier that $\nabla \mathcal{F}^*$ is $\frac{1}{\gamma}$-Lipschitz continuous, therefore:

$$\begin{aligned} \|\nabla \mathcal{D}(\theta) - \nabla \mathcal{D}(\phi)\| &= \|\mathbf{X}(\boldsymbol{s}(\theta) - \boldsymbol{s}(\phi))\| \\ &\leq \|\mathbf{X}\|_2 \|\boldsymbol{s}(\theta) - \boldsymbol{s}(\phi)\| \\ &= \|\mathbf{X}\|_2 \|\nabla \mathcal{F}^*(-U(\theta)) - \nabla \mathcal{F}^*(-U(\phi))\|_F \\ &\leq \frac{\|\mathbf{X}\|_2}{\gamma} \|U(\theta) - U(\phi)\|_F \\ &= \frac{\|\mathbf{X}\|_2}{\gamma} \|\mathbf{X}^T(\theta - \phi)\| \\ &\leq \frac{\|\mathbf{X}\|_2^2}{\gamma} \|\theta - \phi\| \\ &= \frac{\max_i \|\mathbf{X}_i\|_2^2}{\gamma} \|\theta - \phi\|. \end{aligned} \tag{26}$$

Appendix C. Proof of Theorem 6

Let $\mathcal{H}(\boldsymbol{z}, \theta) = \sum_{i=1}^{m} (\|\boldsymbol{z}_i\|_2 - \theta_i^T \boldsymbol{z}_i)$. The Lagrange function in Eq. (5) can be written as

$$\mathcal{L}(\mathbf{W}, \boldsymbol{z}, \theta) = \mathcal{H}(\boldsymbol{z}, \theta) + \mathcal{J}(\mathbf{W}; \theta). \tag{27}$$

Using Algorithm 2, $\mathbf{W}^k$ is a minimizer of $\mathcal{J}_\mu(\mathbf{W}; \theta)$. Then

$$\mathbf{0} \in \frac{\partial}{\partial \mathbf{W}} \mathcal{J}_\mu(\mathbf{W}^k; \theta), \tag{28}$$

where $\frac{\partial}{\partial \mathbf{W}} \mathcal{J}_\mu(\mathbf{W}^k; \theta)$ denotes the sub-gradient of $\mathcal{J}_\mu(\mathbf{W}; \theta)$ w.r.t $\mathbf{W}$ at $\mathbf{W}^k$. Note that $\mathcal{J}_\mu(\mathbf{W}; \theta)$ is strongly convex with parameter γ, and $\mathcal{J}_\mu(\mathbf{W}; \theta)$ is a smooth approximation with parameters $(\gamma + 4\epsilon + \frac{\lambda}{\mu}, \frac{d\lambda}{2})$. For any $\mathbf{W} \in \mathbb{R}^{d \times m}$, we have

$$\begin{aligned} \frac{\gamma}{2} \|\mathbf{W} - \mathbf{W}^k\|_F^2 \leq & \mathcal{J}_\mu(\mathbf{W}; \theta) - \mathcal{J}_\mu(\mathbf{W}^k; \theta) \\ \leq & \mathcal{J}(\mathbf{W}; \theta) + \beta\mu - \mathcal{J}(\mathbf{W}^k; \theta). \end{aligned} \tag{29}$$

Let $\boldsymbol{z}^k = \arg\min_{\boldsymbol{z}} \mathcal{H}(\boldsymbol{z}, \theta^k)$. $\forall \boldsymbol{z} \in \mathbb{R}^{\sum_{i=1}^{m} n_i}$,

$$\mathcal{H}(\boldsymbol{z}, \theta^k) - \mathcal{H}(\boldsymbol{z}^k, \theta^k) \geq 0. \tag{30}$$

Combining Eqs (29) and (30) with (27), we have

$$\frac{\gamma}{2} \|\mathbf{W} - \mathbf{W}^k\|_F^2 \leq \mathcal{L}(\mathbf{W}, \boldsymbol{z}, \theta^k) - \mathcal{L}(\mathbf{W}^k, \boldsymbol{z}^k, \theta^k) + \beta\mu. \tag{31}$$

From Eq. (8), we know that $\mathcal{H}(\boldsymbol{z}^k, \theta^k) = 0$. Therefore

$$\mathcal{D}(\theta^k) = \min_{\mathbf{W}} \mathcal{J}(\mathbf{W}; \theta^k) = \mathcal{J}(\mathbf{W}^k; \theta^k) = \mathcal{L}(\mathbf{W}^k, \boldsymbol{z}^k, \theta^k). \tag{32}$$

Let $\mathbf{W}^*$, $\boldsymbol{z}^*$ and θ^* denote a minimizer of Eq. (5). The equality constraint in (5) implies that $\boldsymbol{z}_i^* = \mathbf{X}_i \boldsymbol{w}_i^* - \boldsymbol{y}_i$. Then we have

$$\mathcal{D}(\theta^*) = \mathcal{L}(\mathbf{W}^*, \boldsymbol{z}^*, \theta^*) = \sum_{i=1}^{m} \|\boldsymbol{z}_i^*\|_2 + \lambda \|\mathbf{W}^*\|_{2,1} + f(\mathbf{W}^*) + g(\mathbf{W}^*) = \mathcal{L}(\mathbf{W}^*, \boldsymbol{z}^*, \theta^k). \tag{33}$$

By combining Eqs (31)-(33) and Theorem 5, we get

$$\frac{\gamma}{2} \|\mathbf{W}^k - \mathbf{W}^*\|_F^2 \leq \mathcal{D}(\theta^*) - \mathcal{D}(\theta^k) + \beta\mu \leq \frac{2\tau L \|\theta^* - \theta^0\|^2}{(k+1)^2} + \beta\mu, \tag{34}$$

where by placing $\beta = \frac{d\lambda}{2}$, we obtain:

$$\|\mathbf{W}^k - \mathbf{W}^*\|_F \leq \sqrt{\frac{4\tau L \|\theta^* - \theta^0\|^2}{\gamma(k+1)^2} + \frac{d\lambda\mu}{\gamma}}. \tag{35}$$

Appendix D. Proof of Theorem 7

For sake of simplicity, let's denote $\mathbf{W}(\theta)$ as $\mathbf{W}$ since it won't cause any confusion. So is $\mathcal{J}(\mathbf{W})$ for $\mathcal{J}(\mathbf{W}; \theta)$. As $\mathcal{J}_\mu$ is $L_f + \frac{\lambda}{\mu} = 4\zeta + \gamma + \frac{\lambda}{\mu}$-smooth, by Theorem 5 we have:

$$\mathcal{J}_\mu(\mathbf{W}^k) - \mathcal{J}_\mu^* \leq 2(L_f + \frac{\lambda}{\mu}) \frac{\|\mathbf{W}_0 - \mathbf{W}_\mu^*\|_F^2}{(k+1)^2}. \tag{36}$$

By making use of Definition 1:

$$\mathcal{J}_\mu(\mathbf{W}) \leq \mathcal{J}(\mathbf{W}) \leq \mathcal{J}_\mu(\mathbf{W}) + \beta\mu. \tag{37}$$

Specifically, we have $\mathcal{J}^* = \mathcal{J}(\mathbf{W}^*) \geq \mathcal{J}_\mu(\mathbf{W}^*) \geq \mathcal{J}_\mu^*$. Combining the above equations we have:

$$\begin{aligned} \mathcal{J}(\mathbf{W}^k) - \mathcal{J}^* &\leq \mathcal{J}_\mu(\mathbf{W}^k) + \beta\mu - \mathcal{J}_\mu^* \\ &\leq 2L_f \frac{\|\mathbf{W}_0 - \mathbf{W}_\mu^*\|_F^2}{k^2} + \frac{2\lambda}{k^2\mu} \|\mathbf{W}_0 - \mathbf{W}_\mu^*\|_F^2 + \beta\mu \end{aligned} \tag{38}$$

Thus, for a certain $K > 0$, for any $k \geq K$:

$$\mathcal{J}(\mathbf{W}^k) - \mathcal{J}^* \leq 2L_f \frac{\|\mathbf{W}_0 - \mathbf{W}_\mu^*\|_F^2}{K^2} + \frac{2\lambda}{K^2\mu} \|\mathbf{W}_0 - \mathbf{W}_\mu^*\|_F^2 + \beta\mu \tag{39}$$

By minimizing the right-hand side regarding μ, we obtain:

$$\mathcal{J}(\mathbf{W}^k) - \mathcal{J}^* \leq 2L_f \frac{\|\mathbf{W}_0 - \mathbf{W}_\mu^*\|_F^2}{K^2} + 2\sqrt{2\lambda\beta} \frac{\|\mathbf{W}_0 - \mathbf{W}_\mu^*\|_F}{K}, \tag{40}$$

where $\mu^* = \frac{\sqrt{2\lambda}\|\mathbf{W}_0-\mathbf{W}^*_\mu\|_F}{K\sqrt{\beta}}$. Therefore, as long as $2L_f\frac{\|\mathbf{W}_0-\mathbf{W}^*_\mu\|_F^2}{K^2}+2\sqrt{2\lambda\beta}\frac{\|\mathbf{W}_0-\mathbf{W}^*_\mu\|_F}{K}\leq\epsilon$, we can guarantee $\mathbf{W}_k$ is an ϵ-optimal solution for $k\geq K$.

Denoting $t=\sqrt{2}\|\mathbf{W}_0-\mathbf{W}^*_\mu\|_F/K$, the above inequality reduces to:

$$L_f t^2+2\sqrt{\lambda\beta}t-\epsilon\leq 0, \tag{41}$$

by the fact that $t>0$, we have:

$$\frac{\sqrt{2}\|\mathbf{W}_0-\mathbf{W}^*_\mu\|_F}{K}=t\leq\frac{-\sqrt{\lambda\beta}+\sqrt{\lambda\beta+L_f\epsilon}}{L_f}=\frac{\epsilon}{\sqrt{\lambda\beta}+\sqrt{\lambda\beta+L_f\epsilon}}, \tag{42}$$

by which we conclude $K\geq\frac{\sqrt{2}\|\mathbf{W}_0-\mathbf{W}^*_\mu\|_F(\sqrt{\lambda\beta}+\sqrt{\lambda\beta+L_f\epsilon})}{\epsilon}$.

If we choose $K=K_1:=\frac{\sqrt{2}\|\mathbf{W}_0-\mathbf{W}^*_\mu\|_F(\sqrt{\lambda\beta}+\sqrt{\lambda\beta+L_f\epsilon})}{\epsilon}$ then we have

$$\begin{aligned}\mu^*&=\frac{\sqrt{2\lambda}\|\mathbf{W}_0-\mathbf{W}^*_\mu\|_F}{K\sqrt{\beta}}=\sqrt{\frac{\lambda}{\beta}}\frac{\epsilon}{\sqrt{\lambda\beta}+\sqrt{\lambda\beta+L_f\epsilon}}\\&=\sqrt{\frac{2}{d}}\frac{\epsilon}{\sqrt{\alpha\beta}+\sqrt{\alpha\beta+L_f\epsilon}},\end{aligned} \tag{43}$$

given $\alpha=\lambda$. Then for any $k\geq K_1$, $\mathcal{J}(\mathbf{W}^k)-\mathcal{J}^*\leq\epsilon$.

On the other hands:

$$\mathcal{J}(\mathbf{W}^*_\mu)-\beta\mu\leq\mathcal{J}_\mu(\mathbf{W}^*_\mu)\leq\mathcal{J}^*\leq\mathcal{J}(\mathbf{W}^0), \tag{44}$$

which along with

$$\mu^*=\sqrt{\frac{\lambda}{\beta}}\frac{\epsilon}{\sqrt{\lambda\beta}+\sqrt{\lambda\beta+L_f\epsilon}}\leq\sqrt{\frac{\lambda}{\beta}}\frac{\epsilon}{\sqrt{\lambda\beta}+\sqrt{\lambda\beta}}\leq\frac{\bar{\epsilon}}{2\beta} \tag{45}$$

implies $\mathcal{J}(\mathbf{W}^*_\mu)\leq\mathcal{J}(\mathbf{W}^0)+\frac{\bar{\epsilon}}{2}$. Since $\mathcal{J}$ has bounded level sets, therefore $\|\mathbf{W}^*_\mu\|_F\leq R_\delta$ where $\delta=\mathcal{J}(\mathbf{W}^0)+\frac{\bar{\epsilon}}{2}$. Thus $\|\mathbf{W}^*_\mu-\mathbf{W}^0\|_F^2\leq(\|\mathbf{W}^*_\mu\|_F+\|\mathbf{W}^0\|_F)^2\leq(R_\delta+\|\mathbf{W}^0\|_F)^2=\Gamma$ and

$$\begin{aligned}K_1&=\frac{\sqrt{2}\|\mathbf{W}_0-\mathbf{W}^*_\mu\|_F(\sqrt{\lambda\beta}+\sqrt{\lambda\beta+L_f\epsilon})}{\epsilon}\\&\leq\frac{\sqrt{2}\|\mathbf{W}_0-\mathbf{W}^*_\mu\|_F(2\sqrt{\lambda\beta}+\sqrt{L_f\epsilon})}{\epsilon}\\&\leq\frac{2\sqrt{2\lambda\beta\Gamma}+\sqrt{2L_f\Gamma\epsilon}}{\epsilon}\\&=\frac{2\lambda\sqrt{d\Gamma}+\sqrt{2L_f\Gamma\epsilon}}{\epsilon}:=K_2,\end{aligned} \tag{46}$$

where we make use of the fact that $\beta=\frac{d\lambda}{2}$, $\sqrt{\gamma+\delta}\leq\sqrt{\gamma}+\sqrt{\delta}$ and $\|\mathbf{W}_0-\mathbf{W}^*_\mu\|_F\leq\sqrt{\Gamma}$.

Therefore, for $k\geq K_2$, we have $\mathcal{J}(\mathbf{W}^k;\theta)-\mathcal{J}(\mathbf{W}^*;\theta)\leq\epsilon$.