Supplemental Material for "Online Robust Non-negative Dictionary Learning for Visual Tracking"

Anonymous ICCV submission

Paper ID 516

1. Equivalence between using Trivial Templates and the Huber Loss Function

Let $\mathbf{R} = \mathbf{Y} - \mathbf{U}\mathbf{V}' = [r_{ij}]$ and $\mathbf{E} = [e_{ij}]$. The objective function $\phi(\mathbf{V}, \mathbf{E}; \mathbf{U})$ in Eqn. 6 of the paper can be expressed as:

$$\begin{aligned}\phi(\mathbf{V}, \mathbf{E}; \mathbf{U}) &= \frac{1}{2}\|\mathbf{Y} - \mathbf{U}\mathbf{V}' - \mathbf{E}\|_F^2 + \lambda\|\mathbf{E}\|_1 + \gamma\|\mathbf{V}\|_1 \\ &= \frac{1}{2}\|\mathbf{R} - \mathbf{E}\|_F^2 + \lambda\|\mathbf{E}\|_1 + \gamma\|\mathbf{V}\|_1 \\ &= \sum_i \sum_j \left[\frac{1}{2}(r_{ij} - e_{ij})^2 + \lambda|e_{ij}|\right] + \gamma\|\mathbf{V}\|_1.\end{aligned} \tag{A1}$$

Since $\phi(\mathbf{V}, \mathbf{E}; \mathbf{U})$ is not differentiable at $e_{ij} = 0$, we cannot use an ordinary gradient method for the minimization w.r.t. e_{ij}. Instead, we resort to a subgradient method. Let us consider the gradients s_+ and s_- to the right and left of $e_{ij} = 0$, respectively:

$$\begin{aligned}s_+ &= \lim_{e_{ij}\to 0^+} \frac{d\phi}{de_{ij}} = -r_{ij} + \lambda \\ s_- &= \lim_{e_{ij}\to 0^-} \frac{d\phi}{de_{ij}} = -r_{ij} - \lambda.\end{aligned} \tag{A2}$$

We consider three cases below:

1. $s_+ \geq 0$ and $s_- \leq 0$ (i.e., $|r_{ij}| \leq \lambda$):

$$e_{ij}^* = 0. \tag{A3}$$

2. $s_+ \leq 0$ and $s_- \leq 0$ (i.e., $r_{ij} \geq \lambda$ and $e_{ij}^* \geq 0$):

$$\frac{d\phi}{de_{ij}} = 0 \Rightarrow -(r_{ij} - e_{ij}) + \lambda = 0 \Rightarrow e_{ij}^* = r_{ij} - \lambda. \tag{A4}$$

3. $s_+ \geq 0$ and $s_- \geq 0$ (i.e., $r_{ij} \leq -\lambda$ and $e_{ij}^* \leq 0$):

$$\frac{d\phi}{de_{ij}} = 0 \Rightarrow -(r_{ij} - e_{ij}) - \lambda = 0 \Rightarrow e_{ij}^* = r_{ij} + \lambda. \tag{A5}$$

These three cases can be summarized by the following which may be regarded as applying a soft-thresholding operation to the residue r_{ij}:

$$e_{ij}^* = \begin{cases} 0 & |r_{ij}| < \lambda \\ \mathrm{sgn}(r_{ij})\,(|r_{ij}| - \lambda) & \text{otherwise.}\end{cases} \tag{A6}$$

We then substitute the optimal e_{ij}^* in Eqn. A6 into $\phi(\mathbf{V}, \mathbf{E}^*; \mathbf{U})$ to eliminate $\mathbf{E}^*$. Two cases are considered separately:

1. If $|r_{ij}| < \lambda$, then $e^*_{ij} = 0$ and hence
$$\frac{1}{2}(r_{ij} - e^*_{ij})^2 + \lambda\,|e^*_{ij}| = \frac{1}{2}r^2_{ij}. \tag{A7}$$

2. If $|r_{ij}| \geq \lambda$, then $e^*_{ij} = \mathrm{sgn}(r_{ij})\,(|r_{ij}| - \lambda)$ and hence
$$\begin{aligned}\frac{1}{2}(r_{ij} - e^*_{ij})^2 + \lambda\,|e^*_{ij}| &= \frac{1}{2}\Big[r_{ij} - \mathrm{sgn}(r_{ij})\,(|r_{ij}| - \lambda)\Big]^2 + \lambda\,(|r_{ij}| - \lambda)\\ &= \frac{1}{2}\Big[|r_{ij}| - (|r_{ij}| - \lambda)\Big]^2 + \lambda(|r_{ij}| - \lambda)\\ &= \frac{1}{2}\lambda^2 + \lambda|r_{ij}| - \lambda^2\\ &= \lambda\,|r_{ij}| - \frac{1}{2}\lambda^2.\end{aligned} \tag{A8}$$

Consequently, we reveal the connections between using trivial templates and the Huber loss function via the objective function in Eqn. 2 of the paper:
$$f(\mathbf{V};\mathbf{U}) = \sum_i \sum_j \ell_\lambda(y_{ij} - \mathbf{u}'_{i\cdot}\mathbf{v}_{j\cdot}) + \gamma\|\mathbf{V}\|_1, \tag{A9}$$
where $\ell_\lambda(\cdot)$ denotes the Huber loss function [2] with parameter λ, which is defined as
$$\ell_\lambda(r) = \begin{cases}\frac{1}{2}r^2 & |r| < \lambda\\ \lambda|r| - \frac{1}{2}\lambda^2 & \text{otherwise.}\end{cases} \tag{A10}$$

2. Proof of Theorem 1

To facilitate proving this theorem, we first define the following surrogate function by expressing the Huber loss as a weighted ℓ_2 loss function:
$$\begin{aligned}h(\mathbf{V};\mathbf{U},\mathbf{W}^p) &= \sum_j h(\mathbf{v}_{j\cdot};\mathbf{U},\mathbf{W}^p)\\ &= \sum_j \left\{\sum_i \left[\frac{1}{2}w^p_{ij}(y_{ij} - \mathbf{u}'_{i\cdot}\mathbf{v}_{j\cdot})^2\right] + \gamma\|\mathbf{v}_{j\cdot}\|_1\right\},\end{aligned} \tag{A11}$$
where
$$w^p_{ij} = \begin{cases}1 & |r^p_{ij}| < \lambda\\ \frac{\lambda}{|r^p_{ij}|} & \text{otherwise.}\end{cases} \tag{A12}$$

The property of this surrogate function and its relationship with the original objective function $f(\mathbf{V};\mathbf{U})$ are given by the two lemmas below:

Lemma 1. *The following inequality holds under the update rule in Eqn. 4 of the paper:*
$$h(\mathbf{v}^{p+1}_{j\cdot};\mathbf{U},\mathbf{W}^p) \leq h(\mathbf{v}^p_{j\cdot};\mathbf{U},\mathbf{W}^p). \tag{A13}$$

Lemma 2. *Based on the definition of* $\mathbf{W}$ *in Eqn. A12, the following inequality holds:*
$$f(\mathbf{V}^{p+1};\mathbf{U}) - f(\mathbf{V}^p;\mathbf{U}) \leq h(\mathbf{V}^{p+1};\mathbf{U},\mathbf{W}^p) - h(\mathbf{V}^p;\mathbf{U},\mathbf{W}^p). \tag{A14}$$

To prove Lemma 1, we first introduce a definition for auxiliary functions:

Definition 1. $g(\mathbf{v}_{j\cdot} \mid \mathbf{v}^p_{j\cdot})$ *is an auxiliary function for* $h(\mathbf{v}_{j\cdot};\mathbf{U})$ *if it satisfies*
$$\begin{aligned}&g(\mathbf{v}_{j\cdot} \mid \mathbf{v}^p_{j\cdot}) \geq h(\mathbf{v}_{j\cdot};\mathbf{U},\mathbf{W}^p),\ \textit{for any } \mathbf{v}_{j\cdot}\\ &g(\mathbf{v}^p_{j\cdot} \mid \mathbf{v}^p_{j\cdot}) = h(\mathbf{v}^p_{j\cdot};\mathbf{U},\mathbf{W}^p).\end{aligned} \tag{A15}$$

Proof of Lemma 1. Let us consider the following auxiliary function $g(\mathbf{v}_{j\cdot} \mid \mathbf{v}^p_{j\cdot})$ for $h(\mathbf{v}_{j\cdot}; \mathbf{U})$:

$$g(\mathbf{v}_{j\cdot} \mid \mathbf{v}^p_{j\cdot}) = h(\mathbf{v}^p_{j\cdot}; \mathbf{U}, \mathbf{W}^p) + (\mathbf{v}_{j\cdot} - \mathbf{v}^p_{j\cdot})' \, \nabla h(\mathbf{v}^p_{j\cdot}; \mathbf{U}, \mathbf{W}^p) + \frac{1}{2}(\mathbf{v}_{j\cdot} - \mathbf{v}^p_{j\cdot})' \, \mathbf{K}(\mathbf{v}^p_{j\cdot}) \, (\mathbf{v}_{j\cdot} - \mathbf{v}^p_{j\cdot}), \tag{A16}$$

where $\mathbf{K}(\mathbf{v}^p_{j\cdot})$ is a diagonal matrix with the (k, k)th diagonal element given by

$$\mathbf{K}(\mathbf{v}^p_{j\cdot})_{kk} = \frac{(\mathbf{U}'\boldsymbol{\Omega}_j \mathbf{U}\mathbf{v}^p_{j\cdot})_k + \gamma}{v^p_{jk}}, \tag{A17}$$

and $\boldsymbol{\Omega}_j$ is a diagonal matrix with the (i, i)th diagonal element being w^p_{ij}. It is obvious that $g(\mathbf{v}^p_{j\cdot} \mid \mathbf{v}^p_{j\cdot}) = h(\mathbf{v}^p_{j\cdot}; \mathbf{U}, \mathbf{W}^p)$. To show that $g(\mathbf{v}_{j\cdot} \mid \mathbf{v}^p_{j\cdot}) \geq h(\mathbf{v}_{j\cdot}; \mathbf{U}, \mathbf{W}^p)$, we first express $h(\mathbf{v}_{j\cdot}; \mathbf{U}, \mathbf{W}^p)$ using the Taylor expansion as follows:

$$h(\mathbf{v}_{j\cdot}; \mathbf{U}, \mathbf{W}^p) = h(\mathbf{v}^p_{j\cdot}; \mathbf{U}, \mathbf{W}^p) + (\mathbf{v}_{j\cdot} - \mathbf{v}^p_{j\cdot})' \, \nabla h(\mathbf{v}^p_{j\cdot}; \mathbf{U}, \mathbf{W}^p) + \frac{1}{2}(\mathbf{v}_{j\cdot} - \mathbf{v}^p_{j\cdot})' \, (\mathbf{U}'\boldsymbol{\Omega}_j \mathbf{U}) \, (\mathbf{v}_{j\cdot} - \mathbf{v}^p_{j\cdot}). \tag{A18}$$

It is sufficient to show that

$$(\mathbf{v}_{j\cdot} - \mathbf{v}^p_{j\cdot})' \left(\mathbf{K}(\mathbf{v}^p_{j\cdot}) - \mathbf{U}'\boldsymbol{\Omega}_j \mathbf{U}\right) (\mathbf{v}_{j\cdot} - \mathbf{v}^p_{j\cdot}) \geq 0. \tag{A19}$$

This is a simple extension of the result in [1]. Next, it is easy to see that each diagonal element of $\mathbf{K}(\mathbf{v}^p_{j\cdot})$ is positive. Thus $\mathbf{K}(\mathbf{v}^p_{j\cdot})$ is a positive definite matrix. This implies that $g(\mathbf{v}_{j\cdot} \mid \mathbf{v}^p_{j\cdot})$ is a strongly convex function with a unique global optimum which is achieved when $\nabla g(\mathbf{v}_{j\cdot} \mid \mathbf{v}^p_{j\cdot}) = 0$. After solving it and rewriting it in matrix form, we obtain the update rule in Eqn. 4 of the paper. □

Proof of Lemma 2 . We first note that because the regularizer $\gamma\|\mathbf{V}\|_1$ cancels out on both sides of the inequality, we can simply omit it in the following proof by focusing only on the loss functions. We consider two cases for each entry:

1. $|r^p_{ij}| < \lambda$:
   In this case, $w^p_{ij} = 1$. By definition, $\ell_\lambda(r^p_{ij}) = \frac{1}{2}(r^p_{ij})^2$. So we have

$$\ell_\lambda(r^{p+1}_{ij}) - \ell_\lambda(r^p_{ij}) = \frac{w^p_{ij}}{2}(r^{p+1}_{ij})^2 - \frac{w^p_{ij}}{2}(r^p_{ij})^2. \tag{A20}$$

2. $|r^p_{ij}| \geq \lambda$:
   In this case, $w^p_{ij} = \lambda/|r^p_{ij}|$. Since

$$\begin{aligned}
\ell_\lambda(r^{p+1}_{ij}) - \ell_\lambda(r^p_{ij}) - \left[\frac{w^p_{ij}}{2}(r^{p+1}_{ij})^2 - \frac{w^p_{ij}}{2}(r^p_{ij})^2\right] &= \lambda|r^{p+1}_{ij}| - \lambda|r^p_{ij}| - \left[\frac{w^p_{ij}}{2}(r^{p+1}_{ij})^2 - \frac{w^p_{ij}}{2}(r^p_{ij})^2\right] \\
&= \lambda|r^{p+1}_{ij}| - \frac{\lambda^2}{w^p_{ij}} - \left[\frac{w^p_{ij}}{2}(r^{p+1}_{ij})^2 - \frac{\lambda^2}{2w^p_{ij}}\right] \\
&= -\frac{w^p_{ij}}{2}\left[(r^{p+1}_{ij})^2 - \frac{2\lambda|r^{p+1}_{ij}|}{w^p_{ij}} + \frac{\lambda^2}{(w^p_{ij})^2}\right] \\
&= -\frac{w^p_{ij}}{2}\left(|r^{p+1}_{ij}| - \frac{\lambda}{w^p_{ij}}\right)^2 \leq 0,
\end{aligned} \tag{A21}$$

   it follows that

$$\ell_\lambda(r^{p+1}_{ij}) - \ell_\lambda(r^p_{ij}) \leq \frac{w^p_{ij}}{2}(r^{p+1}_{ij})^2 - \frac{w^p_{ij}}{2}(r^p_{ij})^2. \tag{A22}$$

By combining Eqn. A20 and Eqn. A22 for all entries, we can prove Lemma 2. □

With these two lemmas, the proof of Theorem 1 in the paper is trivial:

Proof of Theorem 1. From Eqn. A13 and Eqn A14, we get:

$$f(\mathbf{V}^{p+1}; \mathbf{U}) - f(\mathbf{V}^p; \mathbf{U}) \leq h(\mathbf{V}^{p+1}; \mathbf{U}, \mathbf{W}^p) - h(\mathbf{V}^p; \mathbf{U}, \mathbf{W}^p) \leq 0. \tag{A23}$$

So the objective function $f(\mathbf{V}; \mathbf{U})$ is non-increasing under the update rule in Eqn. 4 of the paper. □

3. Proof of Theorem 2

Minimizing $h(\mathbf{V};\mathbf{U},\mathbf{W}^p)$ subject to the non-negativity constraint $\mathbf{V} \geq 0$ is equivalent to minimizing the following Lagrangian function:

$$\mathcal{L}(\mathbf{V};\mathbf{U}) = \sum_j \left\{ \sum_i \left[\frac{1}{2} w_{ij}^p (y_{ij} - \mathbf{u}_{i\cdot}' \mathbf{v}_{j\cdot})^2 \right] + \gamma \|\mathbf{v}_{j\cdot}\|_1 \right\} + \mathrm{tr}(\Phi^T \mathbf{V}), \tag{A24}$$

where Φ denotes the Lagrange multipliers for the non-negativity constraint $\mathbf{V} \geq 0$. Based on the *Karush–Kuhn–Tucker* (KKT) conditions, we have $\Phi_{jk} v_{jk} = 0$. By setting the first derivative of $\mathcal{L}(\mathbf{V};\mathbf{U})$ to 0 for each j, k, we get:

$$-\Big[\sum_i w_{ij}^p y_{ij} \mathbf{u}_{i\cdot}\Big]_k + \Big[\sum_i w_{ij}^p \mathbf{u}_{i\cdot}' \mathbf{u}_{i\cdot} \mathbf{v}_{j\cdot}\Big]_k + \Phi_{jk} + \gamma = 0. \tag{A25}$$

Multiplying both sides by v_{ij} and noting that $\Phi_{jk} v_{jk} = 0$, we get

$$\begin{gathered} \left(-\Big[\sum_i w_{ij}^p y_{ij} \mathbf{u}_{i\cdot}\Big]_k + \Big[\sum_i w_{ij}^p \mathbf{u}_{i\cdot}' \mathbf{u}_{i\cdot} \mathbf{v}_{j\cdot}\Big]_k + \gamma \right) v_{jk} = 0 \\ \Big(-\big[(\mathbf{W}^p \odot \mathbf{Y})'\mathbf{U}\big]_{jk} + \big[\big(\mathbf{W}^p \odot (\mathbf{U}(\mathbf{V}^p)')\big)'\mathbf{U}\big]_{jk} + \gamma \Big) v_{jk} = 0. \end{gathered} \tag{A26}$$

When we apply Eqn. 4 in the paper until convergence, the converged solution satisfies the KKT conditions above. This implies that it is the optimal solution of the original optimization problem.

References

[1] P. Hoyer. Non-negative sparse coding. In *Proceedings of the Workshop on Neural Networks for Signal Processing*, pages 557–565, 2002. 3

[2] P. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. 2