Supplemental Material to Dual-Path Distillation: A Unified Framework to Improve Black-Box Attacks

1 The derivation of Eq. (8)

As described in Section 3.1, the gradient of the efficient-attack loss with respect to the searching direction $\boldsymbol{u}$ (we denote $\boldsymbol{u}_i$ by $\boldsymbol{u}$ for simplicity) can be approximated as

$$\widehat{\boldsymbol{g}}_{\boldsymbol{u}} \approx -\frac{\eta}{q q_v} \sum_{j=1}^{q_v} h\left(\boldsymbol{x}', \boldsymbol{u}, \alpha\right) h\left(\boldsymbol{x}' + \alpha \boldsymbol{u}, \boldsymbol{v}_j, \alpha\beta\right) \boldsymbol{v}_j. \tag{S1}$$

Here, we give the detailed derivation of Eq. (S1) as follows. In this paper, we define two functions $h\left(\boldsymbol{x}', \boldsymbol{u}, \alpha\right) = \frac{f\left(\boldsymbol{x}'+\alpha\boldsymbol{u}\right)-f\left(\boldsymbol{x}'\right)}{\alpha}$ and $\phi\left(\boldsymbol{u}\right) = h\left(\boldsymbol{x}', \boldsymbol{u}, \alpha\right) \boldsymbol{g}_{\boldsymbol{x}}^{\top} \boldsymbol{u}$ which are used in the derivation. According to these definitions, we have

$$\begin{aligned}
\widehat{\boldsymbol{g}}_{\boldsymbol{u}} &= -\frac{\eta}{q q_v} \sum_{j=1}^{q_v} \frac{\phi\left(\boldsymbol{u} + \beta \boldsymbol{v}_j\right) - \phi\left(\boldsymbol{u}\right)}{\beta} \boldsymbol{v}_j \\
&= -\frac{\eta}{q q_v} \sum_{j=1}^{q_v} \left(h\left(\boldsymbol{x}', \boldsymbol{u} + \beta \boldsymbol{v}_j, \alpha\right) \boldsymbol{g}_{\boldsymbol{x}}^{\top} \left(\boldsymbol{u} + \beta \boldsymbol{v}_j\right) - h\left(\boldsymbol{x}', \boldsymbol{u}, \alpha\right) \boldsymbol{g}_{\boldsymbol{x}}^{\top} \boldsymbol{u}\right) \frac{\boldsymbol{v}_j}{\beta} && \text{(S2)} \\
&= -\frac{\eta}{q q_v} \sum_{j=1}^{q_v} \left(h\left(\boldsymbol{x}', \boldsymbol{u} + \beta \boldsymbol{v}_j, \alpha\right) \boldsymbol{g}_{\boldsymbol{x}}^{\top} \boldsymbol{u} - h\left(\boldsymbol{x}', \boldsymbol{u}, \alpha\right) \boldsymbol{g}_{\boldsymbol{x}}^{\top} \boldsymbol{u} + \beta h\left(\boldsymbol{x}', \boldsymbol{u} + \beta \boldsymbol{v}_j, \alpha\right) \boldsymbol{g}_{\boldsymbol{x}}^{\top} \boldsymbol{v}_j\right) \frac{\boldsymbol{v}_j}{\beta} \\
&\approx -\frac{\eta}{q q_v} \sum_{j=1}^{q_v} \boldsymbol{g}_{\boldsymbol{x}}^{\top} \boldsymbol{u} \left(h\left(\boldsymbol{x}', \boldsymbol{u} + \beta \boldsymbol{v}_j, \alpha\right) - h\left(\boldsymbol{x}', \boldsymbol{u}, \alpha\right)\right) \frac{\boldsymbol{v}_j}{\beta} && \text{(S3)} \\
&\approx -\frac{\eta}{q q_v} \sum_{j=1}^{q_v} \boldsymbol{g}_{\boldsymbol{x}}^{\top} \boldsymbol{u} \left(\frac{f\left(\boldsymbol{x}' + \alpha \boldsymbol{u} + \alpha\beta \boldsymbol{v}_j\right) - f\left(\boldsymbol{x}'\right)}{\alpha} - \frac{f\left(\boldsymbol{x}' + \alpha \boldsymbol{u}\right) - f\left(\boldsymbol{x}'\right)}{\alpha}\right) \frac{\boldsymbol{v}_j}{\beta} && \text{(S4)} \\
&\approx -\frac{\eta}{q q_v} \sum_{j=1}^{q_v} \boldsymbol{g}_{\boldsymbol{x}}^{\top} \boldsymbol{u} \left(\frac{f\left(\boldsymbol{x}' + \alpha \boldsymbol{u} + \alpha\beta \boldsymbol{v}_j\right) - f\left(\boldsymbol{x}' + \alpha \boldsymbol{u}\right)}{\alpha\beta}\right) \boldsymbol{v}_j \\
&\approx -\frac{\eta}{q q_v} \sum_{j=1}^{q_v} \boldsymbol{g}_{\boldsymbol{x}}^{\top} \boldsymbol{u} h\left(\boldsymbol{x}' + \alpha \boldsymbol{u}, \boldsymbol{v}_j, \alpha\beta\right) \boldsymbol{v}_j && \text{(S5)} \\
&\approx -\frac{\eta}{q q_v} \sum_{j=1}^{q_v} \frac{f\left(\boldsymbol{x}' + \alpha \boldsymbol{u}\right) - f\left(\boldsymbol{x}'\right)}{\alpha} h\left(\boldsymbol{x}' + \alpha \boldsymbol{u}, \boldsymbol{v}_j, \alpha\beta\right) \boldsymbol{v}_j && \text{(S6)} \\
&\approx -\frac{\eta}{q q_v} \sum_{j=1}^{q_v} h\left(\boldsymbol{x}', \boldsymbol{u}, \alpha\right) h\left(\boldsymbol{x}' + \alpha \boldsymbol{u}, \boldsymbol{v}_j, \alpha\beta\right) \boldsymbol{v}_j.
\end{aligned}$$

Here Eq. (S2) uses the definition of ϕ. And the definition of h is utilized in Eq. (S4) and Eq. (S5). Because $\beta \ll 1$, we neglect $\beta h\left(\boldsymbol{x}', \boldsymbol{u} + \beta \boldsymbol{v}_j, \alpha\right) \boldsymbol{g}_{\boldsymbol{x}}^{\top} \boldsymbol{v}_j$ in Eq. (S3). The term $\boldsymbol{g}_{\boldsymbol{x}}^{\top} \boldsymbol{u}$ in Eq. (S6) is approximated by finite difference method since $\boldsymbol{g}_{\boldsymbol{x}}^{\top} \boldsymbol{u} = D_{\boldsymbol{u}} f\left(\boldsymbol{x}'\right) = \frac{f\left(\boldsymbol{x}'+\alpha\boldsymbol{u}\right)-f\left(\boldsymbol{x}'\right)}{\alpha}$, where $D_{\boldsymbol{u}} f\left(\boldsymbol{x}'\right)$ is the directional derivative of f at a point $\boldsymbol{x}'$ in the direction of a vector $\boldsymbol{u}$.

2 The derivation of Eq. (21)

To simplify the loss Eq. (21) introduced in Section 3.3, we employ the assumption that is also used in [1]. In detail, we assume that all eigenvectors of C have the same eigenvalues, that is, $\mathrm{C} = \sum_{i=1}^{D} \lambda \boldsymbol{p}_i \boldsymbol{p}_i^{\top}$, where $\boldsymbol{p}_i$ is the i^{th} eigenvector.

According to [1], we have $trace\left(\mathrm{C}\right)=1$ and $\|\boldsymbol{p}_i\|_2=1$. It implies that we have $\lambda=\frac{1}{D}$. This yields

$$
\begin{aligned}
\min \ell\left(\widehat{\boldsymbol{g}}_{\boldsymbol{x}}\right) &= -\frac{\left(\boldsymbol{g}_{\boldsymbol{x}}^T \mathrm{C} \boldsymbol{g}_{\boldsymbol{x}}\right)^2}{\left(1-\frac{1}{q}\right) \boldsymbol{g}_{\boldsymbol{x}}^T \mathrm{C}^2 \boldsymbol{g}_{\boldsymbol{x}}+\frac{1}{q} \boldsymbol{g}_{\boldsymbol{x}}^T \mathrm{C} \boldsymbol{g}_{\boldsymbol{x}}} \\
&= -\frac{\left(\boldsymbol{g}_{\boldsymbol{x}}^T \frac{1}{D} \sum_{i=1}^D \boldsymbol{p}_i \boldsymbol{p}_i^\top \boldsymbol{g}_{\boldsymbol{x}}\right)^2}{\left(1-\frac{1}{q}\right) \boldsymbol{g}_{\boldsymbol{x}}^T \frac{1}{D} \sum_{i=1}^D \boldsymbol{p}_i \boldsymbol{p}_i^\top \frac{1}{D} \sum_{j=1}^D \boldsymbol{p}_j \boldsymbol{p}_j^\top \boldsymbol{g}_{\boldsymbol{x}}+\frac{1}{q} \boldsymbol{g}_{\boldsymbol{x}}^T \frac{1}{D} \sum_{i=1}^D \boldsymbol{p}_i \boldsymbol{p}_i^\top \boldsymbol{g}_{\boldsymbol{x}}} \qquad \text{(S7)} \\
&= -\frac{\left(\boldsymbol{g}_{\boldsymbol{x}}^T \frac{1}{D} \sum_{i=1}^D \boldsymbol{p}_i \boldsymbol{p}_i^\top \boldsymbol{g}_{\boldsymbol{x}}\right)^2}{\left(1-\frac{1}{q}\right) \boldsymbol{g}_{\boldsymbol{x}}^T \frac{1}{D^2} \sum_{i=1}^D \sum_{j=1}^D \boldsymbol{p}_i \boldsymbol{p}_i^\top \boldsymbol{p}_j \boldsymbol{p}_j^\top \boldsymbol{g}_{\boldsymbol{x}}+\frac{1}{qD} \boldsymbol{g}_{\boldsymbol{x}}^T \sum_{i=1}^D \boldsymbol{p}_i \boldsymbol{p}_i^\top \boldsymbol{g}_{\boldsymbol{x}}} \\
&= -\frac{\left(\boldsymbol{g}_{\boldsymbol{x}}^T \frac{1}{D} \sum_{i=1}^D \boldsymbol{p}_i \boldsymbol{p}_i^\top \boldsymbol{g}_{\boldsymbol{x}}\right)^2}{\left(1-\frac{1}{q}\right) \boldsymbol{g}_{\boldsymbol{x}}^T \frac{1}{D^2} \sum_{i=1}^D \boldsymbol{p}_i \boldsymbol{p}_i^\top \boldsymbol{g}_{\boldsymbol{x}}+\frac{1}{qD} \boldsymbol{g}_{\boldsymbol{x}}^T \sum_{i=1}^D \boldsymbol{p}_i \boldsymbol{p}_i^\top \boldsymbol{g}_{\boldsymbol{x}}} \qquad \text{(S8)} \\
&= -\frac{\left(\frac{1}{D} \sum_{i=1}^D \boldsymbol{g}_{\boldsymbol{x}}^T \boldsymbol{p}_i \boldsymbol{p}_i^\top \boldsymbol{g}_{\boldsymbol{x}}\right)^2}{\left(1-\frac{1}{q}\right) \frac{1}{D^2} \sum_{i=1}^D \boldsymbol{g}_{\boldsymbol{x}}^T \boldsymbol{p}_i \boldsymbol{p}_i^\top \boldsymbol{g}_{\boldsymbol{x}}+\frac{1}{qD} \sum_{i=1}^D \boldsymbol{g}_{\boldsymbol{x}}^T \boldsymbol{p}_i \boldsymbol{p}_i^\top \boldsymbol{g}_{\boldsymbol{x}}} \\
&= -\frac{\frac{1}{D^2}\left(\sum_{i=1}^D \left(\boldsymbol{g}_{\boldsymbol{x}}^T \boldsymbol{p}_i\right)^2\right)^2}{\left(1-\frac{1}{q}\right) \frac{1}{D^2} \sum_{i=1}^D \left(\boldsymbol{g}_{\boldsymbol{x}}^T \boldsymbol{p}_i\right)^2+\frac{1}{qD} \sum_{i=1}^D \left(\boldsymbol{g}_{\boldsymbol{x}}^T \boldsymbol{p}_i\right)^2} \\
&= -\frac{\frac{1}{D^2}}{\left(1-\frac{1}{q}\right) \frac{1}{D^2}+\frac{1}{qD}} \sum_{i=1}^D \left(\boldsymbol{g}_{\boldsymbol{x}}^T \boldsymbol{p}_i\right)^2 \\
&= -\frac{q}{D+q-1} \sum_{i=1}^D \left(\boldsymbol{g}_{\boldsymbol{x}}^T \boldsymbol{p}_i\right)^2 .
\end{aligned}
$$

Here, we use eigenvectors to represent C in Eq. (S7) and the property, $\boldsymbol{p}_i^\top \boldsymbol{p}_j = \mathbb{1}_{i=j}$, is used in Eq. (S8), where $\mathbb{1}_{i=j}$ is the indicator function.

References

[1] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. *NeurIPS*, 2019.