Different Bregman divergence to be used as the distance generating function in this method. Given a positive definite matrix \( A \), the Mahalanobis norm of a vector \( x \) is defined as \( \| x \|_A = \sqrt{(x, Ax)} \). Let \( g_t = \partial f(s_t) \) be the subgradient of the function being minimized at time \( t \), and \( G_t = \sum_t g_t g_t^T \) be the covariance matrix of outer products of the subgradients. It is computationally more efficient to use the diagonal matrix \( H_t = \sqrt{\text{diag}(G_t)} \) instead of the full covariance matrix, which can be expensive to estimate. Algorithm 3 describes the adaptive subgradient mirror descent TD method.

**Algorithm 3** Composite Mirror Descent TD(\( \lambda \))

1: repeat
2: Do action \( \pi(s_t) \) and observe next state \( s_{t+1} \) and reward \( r_t \).
3: Set TD error \( \delta_t = r_t + \gamma \phi(s_{t+1})^T w_t - \phi(s_t)^T w_t \)
4: Update the eligibility trace \( e_t \leftarrow e_t + \lambda \gamma \phi(s_t) \)
5: Compute TD update \( \xi_t = \delta_t e_t \).
6: Update feature covariance
   \[ G_t = G_{t-1} + \phi(s_t) \phi(s_t)^T \]
7: Compute Mahalanobis matrix \( H_t = \sqrt{\text{diag}(G_t)} \).
8: Update the weights \( w_t \):
   \[ w_{t+1,i} = \text{sign}(w_{t,i} - \frac{\alpha \xi_{t,i}}{H_{t,i}})(w_{t,i} - \frac{\alpha \xi_{t,i}}{H_{t,i}} - \frac{\alpha \beta}{H_{t,i}}) \]
9: Set \( t \leftarrow t + 1 \).
10: until done.

Return \( V^\pi \approx \Phi w_1 \) as the \( l_1 \) penalized sparse value function associated with policy \( \pi \) for MDP \( M \).

**3 Convergence Analysis**

**Definition 2** [GLMH11]: \( \Pi_i \) is the \( l_1 \)-regularized projection defined as: \( \Pi_i y = \Phi \alpha \) such that \( \alpha = \text{arg min}_w \| y - \Phi w \|^2 + \beta \| w \|_1 \), which is a non-expansive mapping w.r.t weighted \( l_2 \) norm induced by the on-policy sample distribution setting, as proven in [GLMH11]. Let the approximation error \( \Phi f(y, \beta) = \| y - \Pi_i y \|^2 \).

**Definition 3** (Empirical \( l_1 \)-regularized projection): \( \hat{\Pi}_i \) is the empirical \( l_1 \)-regularized projection with a specific \( l_1 \) regularization solver, and satisfies the non-expansive mapping property. It can be shown using a direct derivation that \( \hat{\Pi}_i \Pi T \) is a \( \gamma \)-contraction mapping. Any unbiased \( l_1 \) solver which generates intermediate sparse solution before convergence, e.g., SMIDAS solver after \( t \)-th iteration, comprises an empirical \( l_1 \)-regularized projection.

**Theorem 1** The approximation error \( \| V - \hat{V} \| \) of Algorithm 2 is bounded by (ignoring dependence on \( \pi \) for simplicity):

\[
\| V - \hat{V} \| \leq \frac{1}{1 - \gamma} \times \left( \| V - \Pi V \| + f(\Pi V, \beta) + (M - 1) P(0) + \| w^* \|^2 \frac{M}{\alpha N} \right)
\]

where \( \hat{V} \) is the approximated value function after \( N \)-th iteration, i.e., \( \hat{V} = \Phi w_N, M = \frac{2}{\alpha N(1 - \gamma)}, \alpha \) is the step-size, \( P(0) = \frac{1}{N} \sum_{i=1}^{N} \| \Pi V(s_i) \|^2 \), \( s_i \) is the state of \( i \)-th sample, \( e = d^2 \), \( d \) is the number of features, and finally, \( w^* \) is \( l_1 \)-regularized projection of \( \Pi V \) such that \( \Phi w^* = \Pi_i \Pi V \).

**Proof:** In the on-policy setting, the solution given by Algorithm 2 is the fixed point of \( \hat{V} = \hat{\Pi}_i \Pi TV \) and the error decomposition is illustrated in Figure 1. The error can be bounded by the triangle inequality

\[
\| V - \hat{V} \| = \| V - \Pi TV \| + \| \Pi TV - \hat{\Pi}_i \Pi TV \| + \| \hat{\Pi}_i \Pi TV - \hat{V} \|
\]

Since \( \hat{\Pi}_i \Pi T \) is a \( \gamma \)-contraction mapping, and \( \hat{V} = \hat{\Pi}_i \Pi TV \), we have

\[
\| \hat{\Pi}_i \Pi TV - \hat{V} \| = \| \hat{\Pi}_i \Pi TV - \Pi_i \Pi TV \| \leq \gamma \| V - \hat{V} \|
\]

So we have

\[
(1 - \gamma) \| V - \hat{V} \| \leq \| V - \Pi TV \| + \| \Pi TV - \hat{\Pi}_i \Pi TV \|
\]

\[
\| V - \Pi TV \| \text{ depends on the expressiveness of the basis } \Phi, \text{ where if } V \text{ lies in } \text{span}(\Phi), \text{ this error term is zero.}
\]

\[
\| \Pi TV - \hat{\Pi}_i \Pi TV \| \text{ is further bounded by the triangle inequality}
\]

\[
\| \Pi TV - \hat{\Pi}_i \Pi TV \| \leq \| \Pi TV - \hat{\Pi}_i \Pi TV \| + \| \hat{\Pi}_i \Pi TV - \hat{\Pi}_i \Pi TV \|
\]

where \( \| \Pi TV - \hat{\Pi}_i \Pi TV \| \) is controlled by the sparsity parameter \( \beta \), i.e., \( f(\Pi TV, \beta) = \| \Pi TV - \hat{\Pi}_i \Pi TV \| \), where \( \varepsilon = \| \Pi TV - \hat{\Pi}_i \Pi TV \| \) is the approximation error depending on the quality of the \( l_1 \) solver employed. In Algorithm 2, the \( l_1 \) solver is related to the SMIDAS \( l_1 \) regularized mirror-descent method for regression and classification [SST11a]. Note that for a squared loss function...