A Proof of Theorem. 2.1

Proof To prove Theorem. 2.1, we provide an example where the sequence of predictors $\{f_t\}$ is no-regret on loss $\{l_t(f_t)\}$ but Eq. 3 does not hold.

Let us assume there exist a $f^* \in \mathcal{F}$ such that $f^*(x_t) = v_t = \sum_{s=t}^{T} \gamma^{s-t} r_t$. Namely we assume that the best predictor in $\mathcal{F}$ can predict long-term reward exactly. Note that this f^* minimizes the PE and BE simultaneously as $f^* = \arg\min_{f\in\mathcal{F}} \sum (f(\mathbf{x}_t) - v_t)^2$ and $f^* = \arg\min_{f\in\mathcal{F}} \sum l_t(f)$.

Let us assume that $f_t(\mathbf{x}_t) = v_t + a$ and $f_t(\mathbf{x}_{t+1}) = v_{t+1} + \frac{1}{\gamma}a, \forall t$, for $a \in \mathbb{R}^+$. Then we have:

$$b_t = f_t(\mathbf{x}_t) - r_t - \gamma f_t(\mathbf{x_{t+1}}) = v_t + a - r_t - \gamma v_{t+1} - \gamma\frac{1}{\gamma}a = 0. \tag{27}$$

Hence, for regret, we have:

$$\text{Regret} = \sum l_t(f_t) - \min_{f\in\mathcal{F}} \sum l_t(f) = \sum l_t(f_t) - l_t(f^*) = \sum b_t^2 - b_t^{*2} = 0, \tag{28}$$

which means that this sequence of predictor $\{f_t\}$ is no-regret with respect to the loss functions $\{l_t(f)\}$.

However, on the other hand, when we check the predictor error e_t, we have $e_t = f_t(\mathbf{x}_t) - v_t = a$, which makes the sum of prediction error $\sum e_t^2$ increase linearly: $\sum e_t^2 = (T)a^2$ and $(1/T)\sum e_t^2 = a$. Since we have $e_t^* = 0$ for all t and thus $\sum e_t^{*2} = 0$, there is no such constant $C \in \mathbb{R}^+$ that could make the Eq. 3 hold. ∎

This example presents a sequence of predictors that does not satisfy the online stability condition. In fact, it is this example that motivates us to study stability condition of online algorihtms.

B Proof of Lemma. 3.2

Proof Note that $b_t^* = f^*(\mathbf{x}_t) - v_t + v_t - r_t - \gamma f^*(\mathbf{x}_{t+1}) = e_t^* - \gamma e_{t+1}^*$. Squaring both sides and summing over from $t = 0$ to $T - 1$, we have:

$$\begin{aligned}
\sum b_t^{*2} &= \sum (e_t^* - \gamma e_{t+1}^*)^2 \\
&= \sum e_t^{*2} + \sum \gamma^2 e_{t+1}^{*2} - 2\gamma \sum e_t^* e_{t+1}^* \\
&\leq \sum e_t^{*2} + \sum \gamma^2 e_{t+1}^{*2} + \sum \gamma e_t^{*2} + \sum \gamma e_{t+1}^{*2} \\
&= (1+\gamma)^2 \sum e_t^{*2} + (\gamma^2 + \gamma)(e_T^{*2} - e_0^{*2}).
\end{aligned}$$

Again, the first inequality is obtained by applying Young's inequality to $-2e_t^* e_{t+1}^*$ to get $-2e_t^* e_{t+1}^* \leq e_t^{*2} + e_{t+1}^{*2}$. ∎

C Proof of Lemma. 4.1

Proof To show $l_t(f)$ is convex with respect to f, we only need the assumption that $\mathcal{F}$ belongs to a vector space. Since the function space $\mathcal{F}$ belongs to a vector space, for any two function $f \in \mathcal{F}$ and $g \in \mathcal{F}$, and a scalar $a \in R$ and $\mathbf{x}$, we have:

$$(f + g)(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x}), \tag{29}$$

$$(af)(\mathbf{x}) = af(\mathbf{x}). \tag{30}$$

To prove the convexity of the loss functional $l_t(f)$, we show that for any $\alpha \in [0, 1]$, we have $l_t(\alpha f + (1-\alpha)g) \leq \alpha l_t(f) + (1-\alpha)l_t(g)$. For $l_t(\alpha f + (1-\alpha)g)$, we have:

$$l_t(\alpha f + (1-\alpha)g) = ((\alpha f + (1-\alpha)g)(\mathbf{x}_t) - r_t - \gamma(\alpha f + (1-\alpha)g)(\mathbf{x}_{t+1}))^2 \tag{31}$$

$$= (\alpha(f(\mathbf{x}_t) - \gamma f(\mathbf{x}_{t+1}) - r_t) + (1-\alpha)(g(\mathbf{x}_t) - \gamma g(\mathbf{x}_{t+1}) - r_t))^2 \tag{32}$$

$$= \alpha^2(f(\mathbf{x}_t - \gamma f(\mathbf{x}_{t+1}) - r_t))^2 + (1-\alpha)^2(g(\mathbf{x}_t) - \gamma g(\mathbf{x}_{t+1}) - r_t)^2 \tag{33}$$

$$+ 2\alpha(1-\alpha)(f(\mathbf{x}_t) - \gamma f(\mathbf{x}_{t+1}) - r_t)(g(\mathbf{x}_t) - \gamma g(\mathbf{x}_{t+1}) - r_t). \tag{34}$$

For $\alpha l_t(f) + (1-\alpha)l_t(g)$, we have:

$$\alpha l_t(f) + (1-\alpha)l_t(g) = \alpha(f(\mathbf{x}_t) - \gamma f(\mathbf{x}_{t+1}) - r_t)^2 + (1-\alpha)(g(\mathbf{x}_t) - \gamma g(\mathbf{x}_{t+1}) - r_t)^2. \tag{35}$$

Define $b(f) = (f(\mathbf{x}_t) - \gamma f(\mathbf{x}_{t+1}) - r_t)$ and $b(g) = (g(\mathbf{x}_t) - \gamma g(\mathbf{x}_{t+1}) - r_t)$. Subtract Eq. 35 from Eq. 34, we have:

$$l_t(\alpha f + (1-\alpha)g) - (\alpha l_t(f) + (1-\alpha)l_t(g)) \tag{36}$$
$$= (\alpha^2 - \alpha)b(f)^2 + ((1-\alpha)^2 - (1-\alpha))b(g)^2 + 2(\alpha(1-\alpha))b(f)g(f) \tag{37}$$
$$= (\alpha^2 - \alpha)(b(f)^2 + b(g)^2 - 2b(f)g(f)) = (\alpha^2 - \alpha)(b(f) - g(f))^2 \leq 0. \tag{38}$$

Now we prove $l_t(f)$ is Lipschitz continuous. First, consider the case when $\mathcal{F}$ is in RKHS. $l_t(f)$ is differentiable and its gradient is:

$$\nabla l_t(f) = (f(\mathbf{x}_t) - r_t - \gamma f(\mathbf{x}_{t+1}))(K(\mathbf{x}_t, \cdot) - \gamma K(\mathbf{x}_{t+1}, \cdot)). \tag{39}$$

Note that the norm of $\nabla l_t(f)$ is:

$$\|\nabla l_t(f)\| = (f(\mathbf{x}_t) - r_t - \gamma f(\mathbf{x}_{t+1}))^2(1 + \gamma^2 - 2\gamma K(\mathbf{x}_t, \mathbf{x}_{t+1})). \tag{40}$$

Under the assumption that $|f(\mathbf{x})| \leq P$, $|r| \leq R$, $|K(\mathbf{x}_t, \mathbf{x}_{t+1})| \leq K$, it is easy to see that $\|\nabla l_t(f)\|$ is upper bounded by some postive constant. The fact that a function is differentialbe and has bounded gradient implies the function is Lipschitz continuous.

For the case when $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$, we have $l_t(\mathbf{w})$ is differentiable and the gradient is:

$$\nabla l_t(\mathbf{w}) = (\mathbf{w}^T\mathbf{x}_t - r_t - \gamma \mathbf{w}^T\mathbf{x}_t)(\mathbf{x}_t - \gamma \mathbf{x}_{t+1}). \tag{41}$$

Under the assumptions that $\|\mathbf{w}\|_2 \leq W$, $\|\mathbf{x}\|_2 \leq X$, $|r| \leq R$, it is easy to see that $\|\nabla l_t(\mathbf{w})\|_2$ is bounded. Hence, $l_t(\mathbf{w})$ is Lipschitz continuous with respect to L_2 norm.

To see that $l_t(\mathbf{w})$ is also Lipschitz continuous with respect to L_1 norm, note that $\|\nabla l_t(\mathbf{w})\|_\infty$ must be upper bounded, since $|f(\mathbf{x})| \leq P$, $|r| \leq R$, and $|\mathbf{x}^i| \leq X$, where $\mathbf{x}^i$ stands for the i'th entry of the vector $\mathbf{x}$. ∎

D Proof of Lemma. 4.2

Proof Without loss of generality, we assume the regularization $R(f)$ is 1-strongly convex with respect to norm $\|\cdot\|$. Due to strong convexity, we have:

$$\sum_{i=0}^{t} l_i(f_t) + \frac{1}{\mu}R(f_t) \geq \sum_{i=0}^{t} l_t(f_{t+1}) + \frac{1}{\mu}R(f_{t+1}) + \frac{1}{2\mu}\|f_t - f_{t+1}\|. \tag{42}$$

The inequality follows from the fact that $\sum l_t + \frac{1}{\mu}R$ is a strongly convex function and f_{t+1} is a minimizer of $\sum_{i=0}^{t} l_t + \frac{1}{\mu}R$. Similarly, We also have:

$$\sum_{i=0}^{t-1} l_i(f_{t+1}) + \frac{1}{\mu}R(f_{t+1}) \geq \sum_{i=0}^{t-1} l_i(f_t) + \frac{1}{\mu}R(f_t) + \frac{1}{2\mu}\|f_t - f_{t+1}\|, \tag{43}$$

because f_t is a minimizer of $\sum_{i=0}^{t-1} l_i + \frac{1}{\mu_{t-1}}R$. Adding (42) and (43) together side by side and cancelling out repeated terms from both sides, we get:

$$(1/\mu)\|f_t - f_{t+1}\|^2 \leq l_t(f_t) - l_t(f_{t+1}) \leq |l_t(f_t) - l_t(f_{t+1})| \leq L\|f_t - f_{t+1}\| \tag{44}$$

Setting $z = \|f_t - f_{t+1}\|$, and solve the above quadratic inequality with respect to z, we get $\|f_t - f_{t+1}\| \leq L\mu$. Sum from $t = 0$ to T, set $\mu = 1/\sqrt{T}$ and take the limit $T \to \infty$, we get to Eq. 17 ∎

E Proof of Lemma. 4.3

Proof $l_t(\mathbf{w})$ is a quadratic function with respect $\mathbf{w}$. Hence, taking the Tayplor expension of $l_t(\mathbf{w})$ at $\mathbf{w}'$, we have:

$$l_t(\mathbf{w}) = l_t(\mathbf{w}') + \nabla l_t(\mathbf{w}')^T(\mathbf{w}-\mathbf{w}') + \frac{1}{2}(\mathbf{w}-\mathbf{w}')^T\nabla\nabla l_t(\mathbf{w}')(\mathbf{w}-\mathbf{w}'). \tag{45}$$

Note that the Hessian $\nabla\nabla l_t(\mathbf{w}') = 2(\mathbf{x}_t - \gamma\mathbf{x}_{t+1})(\mathbf{x}_t-\gamma\mathbf{x}_{t+1})^T$, which can be writted as:

$$\begin{aligned}\nabla\nabla l_t(\mathbf{w}') &= 2(\mathbf{w}'^T\mathbf{x}_t - r_t - \gamma\mathbf{w}'^T\mathbf{x}_{t+1})^2\frac{(\mathbf{x}_t-\gamma\mathbf{x}_{t+1})(\mathbf{x}_t-\gamma\mathbf{x}_{t+1})^T}{(\mathbf{w}'^T\mathbf{x}_t - r_t-\gamma\mathbf{w}'^T\mathbf{x}_{t+1})^2}\\ &= \frac{1}{2(\mathbf{w}'^T\mathbf{x}_t - r_t - \gamma\mathbf{w}'^T\mathbf{x}_{t+1})^2}\nabla l_t(\mathbf{w}')\nabla l_t(\mathbf{w}')^T \geq \frac{1}{2M}\nabla l_t(\mathbf{w}')\nabla l_t(\mathbf{w}')^T,\end{aligned} \tag{46}$$

where $M = \sup_{\mathbf{w},\mathbf{x}_t,\mathbf{x}_{t+1},r_t}(\mathbf{w}^T\mathbf{x}_t - r_t - \gamma\mathbf{w}^T\mathbf{x}_{t+1})^2$. The derivation in Eq. 46 implicitly assumes that $(\mathbf{w}'^T\mathbf{x}_t - r_t - \gamma\mathbf{w}'^T\mathbf{x}_{t+1}) \neq 0$. But when $(\mathbf{w}'^T\mathbf{x}_t - r_t - \gamma\mathbf{w}'^T\mathbf{x}_{t+1}) = 0$, the final result from Eq. 46 still holds ($\nabla l_t(\mathbf{w}')\nabla l_t(\mathbf{w}')^T = 0$). Note that M is a positive constant since we assume that $\|\mathbf{w}\|$, $\|\mathbf{x}\|$ and $|r_t|$ are all bounded. Hence, we have:

$$l_t(\mathbf{w}) \geq l_t(\mathbf{w}') + \nabla l_t(\mathbf{w}')^T(\mathbf{w}-\mathbf{w}') + \frac{1}{4M}(\mathbf{w}-\mathbf{w}')^T\nabla l_t(\mathbf{w}')\nabla l_t(\mathbf{w}')^T(\mathbf{w}-\mathbf{w}'). \tag{47}$$

Setting $\lambda \leq 1/2M$ we prove the lemma. ∎

F Proof of Lemma. 4.4

Proof We next show that online newton method satisfies the online stability condition. For convenience, define $\mathbf{y}_{t+1} = \mathbf{w}_t - \frac{1}{\lambda}A_t^{-1}\nabla l_t(\mathbf{w}_t)$, we have:

$$\begin{aligned}&\sum\|\mathbf{w}_t - \mathbf{w}_{t+1}\|_{A_t}^2 \leq \sum\|\mathbf{w}_t - \mathbf{y}_{t+1}\|_{A_t}^2 = \sum\|\frac{1}{\lambda}A_t^{-1}\nabla l_t(\mathbf{w}_t)\|_{A_t}^2\\ &= \sum\frac{1}{\lambda^2}\nabla l_t(\mathbf{w}_t)^TA_t^{-1}A_tA_t^{-1}\nabla l_t(\mathbf{w}_t) = \frac{1}{\lambda^2}\sum\nabla l_t(\mathbf{w}_t)^TA_t^{-1}\nabla l_t(\mathbf{w}_t).\end{aligned}$$

Following the proof in Hazan et al. (2006), it can be shown that:

$$\sum\nabla l_t(\mathbf{w}_t)^TA_t^{-1}\nabla l_t(\mathbf{w}_t) \leq n\log(\frac{TG^2}{\epsilon}+1),$$

where $G\in\mathbb{R}^+$ and $G \geq \|\nabla l_t\|_2$. We simply set $\epsilon = G^2$. Hence, the online stability condition is satisfied as:

$$\begin{aligned}&\frac{1}{T}\sum(\mathbf{w}_t^T\mathbf{x}_{t+1} - \mathbf{w}_{t+1}^T\mathbf{x}_{t+1})^2 \leq \frac{X^2}{T}\sum\|\mathbf{w}_t - \mathbf{w}_{t+1}\|_2^2\\ &\leq \frac{1}{T}\frac{X^2}{\epsilon}\sum\|\mathbf{w}_t - \mathbf{w}_{t+1}\|_{A_t}^2 \leq \frac{1}{T}\frac{X^2}{G^2\lambda^2}n\log(T+1) = 0, \quad T\to\infty.\end{aligned}$$

The first inequality comes from Cauchy-Schwarz inequality and the assumption that $\|\mathbf{x}\|_2 \leq X$. The second inequality follows from the fact that the smallest eigenvalues of A_t's are bigger than or equal to ϵ. ∎

G Disccusion and Results for Online Algorithms on TD-loss Functions $\{\tilde{l}_t(f)\}$

First of all, we show similar to our analysis for Bellman Residual minimization, simply being no-regret on the TD-loss functions $\{\tilde{l}_t(f)\}$ under general function approximation (not necessarily linear) is not sufficient to small predictive errors (Eq. 3 doesn't hold):

Theorem G.1 *There exists a sequence of $\{f_t\}$ that is no-regret with respect to the TD-loss functions $\{\tilde{l}_t(f)\}$, but no $C\in\mathbb{R}^+$ exists that makes Eq. 3 hold.*

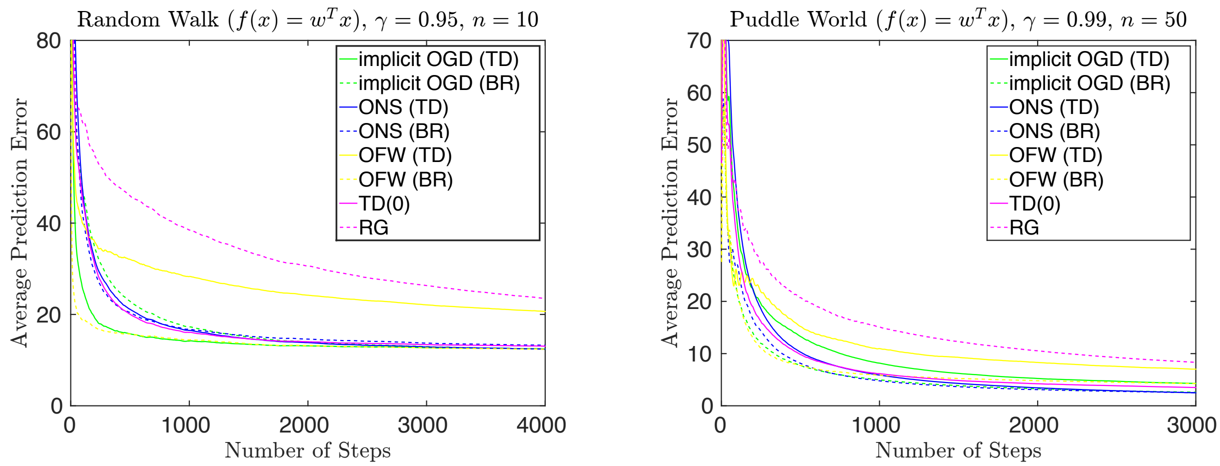

Figure 4: Convergence of prediction error. We applied a set of online algorithms (OGD, implicit OGD, ONS, OFW) on BE loss functions $\{l_t(\mathbf{w})\}$ (dot line) and TD-loss functions $\{\tilde{l}_t(\mathbf{w})\}$ (solid line) for Random walk (left) and Puddle World (right).

Proof Again, we assume that $f_t(\mathbf{x}_t) = v_t + a$ and $f_t(\mathbf{x}_{t+1}) = v_{t+1} + \frac{1}{\gamma}a$, where $a \in \mathbb{R}^+$ and $v_t = \sum_{s=t}^{T} \gamma^{s-t} r_t$ is the long-term reward. Under this setting, the TD-loss $\tilde{l}_t(f_t)$ becomes:

$$\tilde{l}_t(f_t) = (f_t(\mathbf{x}_t) - r_t - \gamma f_t(\mathbf{x}_{t+1}))^2 = 0. \tag{48}$$

Hence, this sequence of predictors $\{f_t\}$ is no-regret:

$$\sum \tilde{l}_t(f_t) - \sum \tilde{l}_t(f^*) \leq \sum \tilde{l}_t(f_t) = 0, \quad \forall f^* \in \mathcal{F}. \tag{49}$$

However this sequence of predictors performly badly in terms of prediction error $e_t^2 = (f_t(\mathbf{x}_t) - v_t)^2 = a^2$. Under the assumption that the function space $\mathcal{F}$ (hypothesis class) is broad enough to have f^* that perfectly predicts long-term reward ($f^*(\mathbf{x}_t) = v_t, \forall t$), we always have $\frac{1}{T}\sum e_t^2 = a^2 > \frac{1}{T} e_t^{*2} = 0$. Hence, it is impossible to find a postive constant C such that Eq. 3 will hold. ∎

Note that in the above proof, the constructed predictors are not stable in a sense that $f_t(\mathbf{x}_{t+1})$ and $f_{t+1}(\mathbf{x}_{t+1})$ varies a lot and hence it does not satisfies the online stability condition.

We conjecture that together with a similar stability analysis as we did for Bellman Residual minimization, we could achieve similar predictive guarantees as in Theorem. 3.3. We leave it as a open question here and we currently are working on it.

G.1 Empirical Results

We applied several stable no-regret online learning algorithms including ONS, OFW, implicit OGD to TD-loss functions $\tilde{l}_t(f)$ with linear function approximation ($f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$). Fig. 4 shows the results of applying the set of algorithms (OGD, implicit OGD, ONS, and OFW) to BE and TD-loss for Random Walk and Puddle World. We compare their performance to $TD(0)$ and $RG(0)$. Although we currently do not have sound predictive guarantees, these empirical results suggest that applying stable no-regret online algorithms to TD-loss functions $\{\tilde{l}_t(\mathbf{w})\}$ in practice may give competitive performance compared to Bellman Residual minimization algorithms and the $TD(0)$.