Theory of Machine Learning

Tommy He

1 General ML Bounds

Definition 1.1. Some initial commonly used definitions and notations:

- *Domain*: set of objects we are trying to label; denoted $\mathcal{X}$ with *instances* $x \in \mathcal{X}$

- *Labels*: set of labels for the objects in the domain; denoted $\mathcal{Y}$ with $y \in \mathcal{Y}$

- *Data generation model*: an unknown, arbitrary data distribution; denoted $\mathcal{D}, \rho$

- *Training data*: finite sequence of pairs in $\mathcal{X} \times \mathcal{Y}$ drawn iid from $\mathcal{D}$; denoted $\mathcal{S} = \mathcal{S}_m = \{(x_1, y_1), \cdots, (x_m, y_m)\}$

- *Hypothesis/predictor/classifier*: rule used by the learner to predict the label of new domain points; denoted $h : \mathcal{X} \to \mathcal{Y}$ with $h \in$ *hypothesis class* $\mathcal{H}$

- *Concept*: map from domain to labels which we are trying to learn; denoted $C : \mathcal{X} \to \mathcal{Y}$ with $c \in$ *concept class* C

- *True labels*: function on $\mathcal{X}$ giving the true labels for each domain point; denoted $f : \mathcal{X} \to \mathcal{Y}$

- *Loss*: an arbitrary loss function on between predicted value and true value; denoted $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$

- *0-1 loss*: loss on $\mathcal{Y} \times \mathcal{Y}$ defined by $(y, y') \mapsto \begin{cases} 0 & \text{if } y = y' \\ 1 & \text{else} \end{cases}$; denoted l_{0-1}

- *Global/generalization loss/error of predictor/risk*: the generalization error of a hypothesis $h \in \mathcal{H}$ i.e. the expected error on a distribution $\mathcal{D}$; denoted $L, R, L_{(\mathcal{D},f)}$.

$$L(h) = L_{(\mathcal{D},f)}(h) = \mathbb{P}_{x \sim \mathcal{D}}(h(x) \neq f(x)) = \mathbb{E}_{x \sim \mathcal{D}}[l_{0-1}(h(x), f(x))]$$

- *Empirical/training error/risk*: the average error of a hypothesis $h \in \mathcal{H}$ over sample from $\mathcal{D}$; denoted $L_\mathcal{S}, \widehat{L}_\mathcal{S}, \widehat{R}_\mathcal{S}, \widehat{R}$

$$\widehat{L}_\mathcal{S}(h) = \frac{1}{m} \sum_{i=1}^{m} l_{0-1}(h(x_i, y_i))$$

We often choose h to minimize $L_\mathcal{S}(h)$, the empirical error, when we actually want to minimize $L(h)$, the generalization error. So, we wish to study the gap $L(h) - L_\mathcal{S}(h)$ to see what we can say about it.

Definition 1.2. Concept class C is *PAC-learnable (Probably Approximately Correct)* if $\forall c \in C, \exists$ learning algo. s.t. $\forall \varepsilon, \delta > 0, \exists m = f(\varepsilon, \delta)$ for function f s.t. if you have m iid samples on any distr. $\mathcal{D}$ on $\mathcal{X}$,

$$\mathbb{P}_{S \sim \mathcal{D}^m}(L_{\mathcal{D}}(h_S) \leq \varepsilon) \geq 1 - \delta$$

Definition 1.3. $\mathcal{H}$ is *realizable*, if $\exists$ target concept $c \in \mathcal{H}$ with $L(c) = 0$. Hypothesis h_S is *consistent* over S if $\hat{L}_S(h_S) = 0$. We have $y_i = c(x_i), h_S(x_i)$ for $1 \leq i \leq m$.

Theorem 1.4 (Learning Bounds; finite $\mathcal{H}$, consistent). *Let $\mathcal{H}$ be finite and realizable with target $c \in \mathcal{H}$, sample S, and algo. $\mathcal{A}$ returning consistent hypothesis $h_S = \mathcal{A}(S, c)$. Then, $\forall \varepsilon, \delta > 0$,*

$$\mathbb{P}_{S \sim \mathcal{D}^m}(L(h_S) \leq \varepsilon) \geq 1 - \delta \text{ if } m \geq \frac{1}{\varepsilon}\left(\log \mathcal{H} + \log \frac{1}{\delta}\right)$$

$$\Leftrightarrow L(h_S) \leq \frac{1}{m}\left(\log \mathcal{H} + \log \frac{1}{\delta}\right) \text{ with prob. } \geq 1 - \delta$$

Proof. For $\varepsilon > 0$, let $\mathcal{H}_\varepsilon := \{h \in \mathcal{H} \mid R(h) > \varepsilon\}$. Then for $h \in \mathcal{H}_\varepsilon$, $\qquad\qquad\qquad\qquad\qquad$ □

Lemma 1.5 (Hoeffding's). *For iid rv $\theta_1, \cdots, \theta_n$ with $\mathbb{E}\theta_i = \mu$ and $\mathbb{P}(a \leq \theta_i \leq b) = 1 \forall i : 1 \leq i \leq n$, we have $\forall \varepsilon > 0$,*

$$\mathbb{P}\left(*\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu > \varepsilon\right) \leq 2\exp\left(\frac{-2m\varepsilon^2}{(b-a)^2}\right)$$

Corollary 1.6.
$$\mathbb{P}(L_{S_{test}}(h) - L_{\mathcal{D}}(h) \geq \varepsilon) \leq 2\exp(-2m\varepsilon^2)$$

where the loss function is in $[0, 1]$ and S_{test} denotes a test set. Note that this no longer applies post-training as the loss functions are no longer independent.

Proof. Apply Hoeffding's Lemma to the test set $S_{\text{test}} = \{(x_1, y_1, \cdots, (x_m, y_m)\}$, and note that $L_{S_{\text{test}}} = \frac{1}{m}\sum_{i=1}^{m} l_{0-1}(h(x_i), y_i)$ and (x_1, y_1) are iid from $\mathcal{D}$. $\qquad\qquad\qquad\qquad$ □

Corollary 1.7. $\forall \delta > 0$,

$$L_{S_{test}}(h) - L_{\mathcal{D}}(h) \leq \sqrt{\frac{\log(2/\delta)}{2m}} \text{ with prob. } \geq 1 - \delta$$

Proof. Since $\mathbb{P}(L_S(h) - L_{\mathcal{D}}(h) \geq \varepsilon) \leq 2\exp(-2m\varepsilon^2)$, we have $\mathbb{P}(L_S(h) - L_{\mathcal{D}}(h) < \varepsilon) > 1 - 2\exp(-2m\varepsilon^2)$, which means we can write

$$\mathbb{P}(L_S(h) - L_{\mathcal{D}}(h) \leq \varepsilon) \geq \mathbb{P}(L_S(h) - L_{\mathcal{D}}(h) < \varepsilon)$$
$$\geq 1 - 2\exp(-2m\varepsilon^2)$$

Now let us substitute $\varepsilon = \sqrt{\frac{\log(2/\delta)}{2m}}$. Solving this also gives $2m\varepsilon^2 = \log\frac{2}{\delta}$. Plugging in,

$$\mathbb{P}\left(L_S(h) - L_{\mathcal{D}}(h) \leq \sqrt{\frac{\log\frac{2}{\delta}}{2m}}\right) \geq 1 - 2\exp(-2m\varepsilon^2)$$

$$= 1 - 2\exp\left(-\log\frac{\delta}{2}\right)$$

$$= 1 - \delta$$

as desired. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ □

Rademacher Complexity

Definition 1.8. A *Rademacher variable* is an rv $\sigma = (\sigma_1, \cdots, \sigma_m)$ for iid rd $\theta_i \forall i : 1 \leq i \leq m$ with uniform values in $\{-1, 1\}$.

Notation 1.9. Let $\mathcal{G}$ be the family of loss functions associated with $\mathcal{H}$ for an arbitary loss $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Explicity,

$$\mathcal{G} = \{g \text{ defined by } (x, y) \mapsto l(h(x), y) \mid h \in \mathcal{H}\}; g \in \mathcal{G}$$

Let $Z = X \times Y$ where $z_i = (x_i, y_i) \forall i : 1 \leq i \leq m$.

Definition 1.10. The *empirical Rademacher complexity* for $S = (z_1, \cdots, z_m)$ is

$$\widehat{\mathcal{R}}_S(\mathcal{G}) = \frac{1}{m} \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \sum_{i=1}^m \sigma_i g(\mathcal{S}_i) \right]$$

Definition 1.11. The *Rademacher complexity* for distrbution $\mathcal{D}, m \geq 1$ is

$$\mathcal{R}_m(\mathcal{G}) = \mathbb{E}_{S \sim \mathcal{D}^m} \widehat{\mathcal{R}}_S(\mathcal{G})$$

Definition 1.12. Function $\Phi : Z^m \to \mathbb{R}$ satisfies *bounded difference inequality* with constant $c = (c_1, \cdots, c_m)$ if $\forall z_1, \cdots, z_m, z_j' \in Z, j : 1 \leq j \leq m$,

$$\Phi(z_1, \cdots, z_j, \cdots, z_m) - \Phi(z_1, \cdots z_j', \cdots, z_m) \leq c_j$$

Theorem 1.13 (McDiarmid's). *Let* $\mathcal{S} = (z_1, \cdots, z_m)$ *for iid rv* $z_i \forall i : 1 \leq i \leq m$ *and* $\Phi : Z^m \to \mathbb{R}$ *satisfy bounded difference inequality with* $c = (c_1, \cdots, c_m)$. *Then*

$$\mathbb{P}(\Phi(\mathcal{S}) - \mathbb{E}\Phi(\mathcal{S}) \geq \varepsilon) \leq 2 \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^m c_i^2} \right)$$

$\forall \varepsilon > 0$.

Example 1.14. Take $\Phi : Z^m \to \mathbb{R}$ defined by $\Phi(\mathcal{S}) = \frac{1}{m} \sum_{i=1}^m x_i = \widehat{\mu}$ where $x_i \in [-a, a] \forall i : 1 \leq i \leq m$. Then Φ satisfies bounded difference inequality with $c = (\frac{2a}{m}, \cdots, \frac{2a}{m})$, which we can check

$$\Phi(\mathcal{S}) - \Phi(\mathcal{S}') = *\frac{1}{m}(x_i - x_i') \leq \frac{x_i - x_i'}{m} \leq \frac{2a}{m}$$

By McDiarmid's Theorem, we can conclude

$$\mathbb{P}(\widehat{\mu} - \mu \geq \varepsilon) \leq \exp\left(-\frac{m\varepsilon^2}{a^2} \right)$$

Notation 1.15. Let $\widehat{\mathbb{E}}_\mathcal{S} g$ be the empirical loss with g given by

$$\widehat{\mathbb{E}}_\mathcal{S} g = \frac{1}{m} \sum_{i=1}^m g(z_i)$$

Let $(h)^+$ denote $\max(h, 0)$.

Theorem 1.16. *Let $\mathcal{G}$ be the losses mapping into $[0, 1]$. Then $\forall g \in \mathcal{G}, \delta > 0$,*

$$\mathbb{E}g(z) \le \widehat{\underset{S}{\mathbb{E}}}g + 2\mathcal{R}_m(\mathcal{G}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \text{ with prob. } \ge 1 - \delta$$

Proof. Let Φ defined by $\Phi(S) = \sup_{g \in \mathcal{G}} \left(\mathbb{E}g - \widehat{\underset{S}{\mathbb{E}}}g \right)$. We find that

$$\Phi(S) - \Phi(S') = {}_{*}\sup_{g \in \mathcal{G}} \left(\mathbb{E}g - \widehat{\underset{S}{\mathbb{E}}}g \right) - \sup_{g \in \mathcal{G}} \left(\mathbb{E}g - \widehat{\underset{S'}{\mathbb{E}}}g \right) \tag{1}$$

$$= {}_{*}\sup_{g \in \mathcal{G}} \left(\widehat{\underset{S}{\mathbb{E}}}g - \widehat{\underset{S'}{\mathbb{E}}}g \right) \tag{2}$$

$$= {}_{*}\sup_{g \in \mathcal{G}} \left(\frac{g(z_i) - g(z_i')}{m} \right) \tag{3}$$

$$\le \frac{1}{m} \tag{4}$$

where (2) comes from knowing $\mathbb{E}g - \widehat{\underset{S}{\mathbb{E}}}g$ and $\mathbb{E}g - \widehat{\underset{S}{\mathbb{E}}}g$ are bounded, since each $g \in \mathcal{G}$ maps into $[0, 1]$. Similarly for (4), this implies $-1 \le g(z_i) - g(z_i') \le 1$ to finish. With this, we know Φ satisfies the bounded inequality with $(\frac{1}{m}, \cdots, \frac{1}{m})$, which allows us to apply McDiarmid's Theorem to give

$$\mathbb{P}(\Phi(S) - \mathbb{E}\Phi(S) \ge \varepsilon) \le 2\exp(-2m\varepsilon^2)$$

$\forall \varepsilon > 0$, which we can rewrite with $\varepsilon = \sqrt{\frac{\log(2/\delta)}{2m}}$ to give

$$\Phi(S) \le \underset{S}{\mathbb{E}}\Phi(S) + \sqrt{\frac{\log\frac{2}{\delta}}{2m}} \text{ with prob. } \ge 1 - \frac{\delta}{2}$$

$\forall \delta > 0$. Replacing $\frac{\delta}{2}$ with δ now gives

$$\Phi(S) \le \underset{S}{\mathbb{E}}\Phi(S) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \text{ with prob. } \ge 1 - \delta$$

Notice we also have

$$\underset{S}{\mathbb{E}}\Phi(S) = \underset{S}{\mathbb{E}}\left[\sup_{g \in \mathcal{G}} \left(\mathbb{E}g - \widehat{\underset{S}{\mathbb{E}}}g \right) \right]$$

By definition, $\mathbb{E}g = \underset{\mathcal{S}'}{\mathbb{E}} \underset{\mathcal{S}'}{\widehat{\mathbb{E}}} g$, so we can rewrite the above as

$$\underset{\mathcal{S}}{\mathbb{E}} \left[\sup_{g \in \mathcal{G}} \left(\underset{\mathcal{S}'}{\mathbb{E}} \left[\underset{\mathcal{S}'}{\widehat{\mathbb{E}}} g - \underset{\mathcal{S}}{\widehat{\mathbb{E}}} g \right] \right) \right] \leq \underset{\mathcal{S},\mathcal{S}'}{\mathbb{E}} \left[\sup_{g \in \mathcal{G}} \left(\underset{\mathcal{S}'}{\widehat{\mathbb{E}}} g - \underset{\mathcal{S}}{\widehat{\mathbb{E}}} g \right) \right] \tag{5}$$

$$= \underset{\mathcal{S},\mathcal{S}'}{\mathbb{E}} \left[\sup_{g \in \mathcal{G}} \left(\frac{1}{m} \sum_{i=1}^{m} (g(z_i') - g(z_i)) \right) \right] \tag{6}$$

$$= \underset{\sigma,\mathcal{S},\mathcal{S}'}{\mathbb{E}} \left[\sup_{g \in \mathcal{G}} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i (g(z_i') - g(z_i)) \right) \right] \tag{7}$$

$$= \underset{\sigma,\mathcal{S}}{\mathbb{E}} \left[\sup_{g \in \mathcal{G}} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i (g(z_i)) \right) \right] + \underset{\sigma,\mathcal{S}'}{\mathbb{E}} \left[\sup_{g \in \mathcal{G}} \left(\frac{1}{m} \sum_{i=1}^{m} -\sigma_i (g(z_i')) \right) \right] \tag{8}$$

$$= 2\mathcal{R}_m(\mathcal{G}) \tag{9}$$

where the Rademacher variables in (7) hold beacause if $\sigma_i = 1$, then the equation is the same as before. If $\sigma_i = 1$, then we can flip that particular z_i and z_i' to get the same. Since it is the expetation over all $\mathcal{S}, \mathcal{S}'$, equality still holds. (8) comes from the linearity of sup and $\mathbb{E}$. Putting the results together yields the desired equation. $\qquad \square$

VC Dimension

Definition 1.17. Let $\mathcal{H}$ be a hypothesis class and finite $C \subseteq \mathcal{X}$. Then $\mathcal{H}$ *shatters* C if $\mathcal{H}$ realizes all possible labels on C.

Definition 1.18. The VC dimension of a hypothesis class $\mathcal{H}$, denoted VCdim($\mathcal{H}$), is the size of the largest C s.t. $\mathcal{H}$ shatters C.

Example 1.19. Let $\mathcal{X} = \mathbb{R}^2, \mathcal{Y} = \{0, 1\}, \mathcal{H} = \{h : \mathbb{R}^2 \to \{0, 1\}, h(x_1, x_2) = \text{sgn}(\beta_2 x_2 + \beta_1 x_1 + \beta_0)\}$. Note that as long as we can find some C with size k s.t. $\mathcal{H}$ shatters C, then VCdim($\mathcal{H}$) $\geq k$. In this case VCdim($\mathcal{H}$) = 4 as we can always find an assignment in $\{0, 1\}$ for the 4 points s.t. the line does not separate them correctly. In general, in $\mathbb{R}^d$, VCdim($\mathcal{H}$) = $d + 1$.

Example 1.20. For fixed $k \in \mathbb{N}$, let $\mathcal{X} = \mathbb{R}, \mathcal{Y} = \{0, 1\}, \mathcal{H} = \{h : \mathbb{R} \to \{0, 1\}, h = \mathbb{1}(\text{union of } k \text{ intervals})\}$. Here, we have VCdim($\mathcal{H}$) = $2k$ by looking at the case where we alternate the points between 0, 1.

Example 1.21. Let $\mathcal{X} = \mathbb{R}^2, \mathcal{Y} = \{0, 1\}, \mathcal{H} = \{h : \mathbb{R}^2 \to \{0, 1\}, h = \mathbb{1}(\text{axis-aligned rectangle})\}$. Then, VCdim($\mathcal{H}$) = 4 since if we take any 5 points and label the outer four with 1 and inner one with 0, then it can not be achieved.

Theorem 1.22 (Vapnik)**.**

$$R \leq \widehat{R} + \sqrt{\frac{1}{m} \left(H + H \log \left(\frac{2m}{H} \right) - \log \left(\frac{\delta}{4} \right) \right)} \quad \textit{with prob.} \ \geq 1 - \delta$$

where H is the VC dimension.

2 Reproducing Hilbert Kernel Spaces (RHKS)

Definition 2.1. Function $K : X \times X \to \mathbb{R}$ is an *SPSD kernel* if

1. K is symmetric (S) i.e. $K(x, y) = K(y, x) \forall x, y \in X$

2. K is positive semi-definite i.e. $\forall x_1, \cdots, x_n \in X, [K_{ij}] = [K(x_i, x_j)]$ is a positive semi-definite (PSD) matrix.

Example 2.2. For $X = \mathbb{R}^d, K : X \times X \to \mathbb{R}$ defined by $K(x, z) = (\langle x, z \rangle)^m$ or $K(x, z) = (1 + \langle x, z \rangle)^m$ for $m \geq 2$ is an SPSD kernel, specifically the *polynomial kernel*.

Example 2.3. $\phi_i = \sin\left(\frac{(2i-1)\pi x}{2}\right)$ for $i = 1, 2, \cdots$ is an orthonormal basis for functions $[0, 1] \to \mathbb{R}$. Take the function vector defined by $\Phi = (\sqrt{\mu_1}\phi_1, \sqrt{\mu_2}\phi_2, \cdots, \sqrt{\mu_i}\phi_i, \cdots)$ with $\sum_{i=0}^{\infty} \mu_i < \infty, \mu_i \geq 0 \forall i$ (where the μ_i are to shrink the norm to be finite). Then, the function $K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ defined by $(x, z) \mapsto \langle \phi(x), \phi(z) \rangle$ is an SPSD kernel.

Definition 2.4. Hilbert space H of functions on set X is a *reproducing kernel hilbert space (RKHS)* if the evaluation functional over H is continuous i.e. $L_x := f \mapsto f(x) \forall f \in H, x \in X$ is continuous or bounded i.e. $\exists M_x > 0$ s.t. $|L_x(f)| = |f(x)| \leq M_x \|f\|_H \forall f \in H, x \in X$.

Proposition 2.5. *By the Riesz representation theorem, $\forall x \in X, \exists! K_x \in H$ with the reproducing property s.t. $\forall f \in H, L_x(f) = \langle f, K_x \rangle_H$. By Riesz again, for $y \in X, \exists! K_y \in H$ with $K_x(y) = L_y(K_x) = \langle K_x, K_y \rangle_H$.*

Definition 2.6. The *reproducing kernel* of hilbert space H is defined by $K : X \times Y \to \mathbb{R}, (x, y) \mapsto \langle K_x, K_y \rangle_H$, which is an SPSD kernel, for K_x, K_y defined as above

Theorem 2.7 (Moore-Aronszajn). *Every SPSD kernel K on X defines uniquely a RKHS of functions on X for which K is a reproducing kernel.*

Theorem 2.8 (Representer Theorem). *Take a SPSD kernel $K : X \times X \to \mathbb{R}$ with RKHS H. Let $(x_1, y_1), \cdots, (x_n, y_n) \in X \times \mathbb{R}$ be a training sample, $g : [0, \infty) \to \mathbb{R}$ a strictly increasing function, and $E : (X \times \mathbb{R})^n \to \mathbb{R} \cup \{\infty\}$ be an error function. Any minimizer of the regularized empirical risk functional $f \mapsto E((x_1, y_1, f(x_1)), \cdots, (x_n, y_n, f(x_n))) + g(\|f\|)$ on H has the form $x \mapsto \sum_{i=1}^{n} \alpha_i K(x, x_i)$ for $\alpha_i \in \mathbb{R} \forall i$.*

Remark 2.9. The representer theorem greatly reduces the complexity required to search for an optimal function as it reduces the search space from all functions in $|H|$ to the n constants α_i in $\mathbb{R}^n$.

3 Reinforcement Learning

Definition 3.1. The *probability simplex* of a set N is the set of all possible probablity distributions over N, denoted $\Delta(N)$. If N is discrete, then

$$\Delta(N) = \left\{ x \in \mathbb{R}_{\geq 0}^{|N|} \,\middle|\, \sum_{n=1}^{|N|} x_n = 1 \right\}$$

Definition 3.2. A *Markov Decision Process (MDP)* is a tuple $M = (S, A, P, r, \gamma, \mu)$ where

- S is the *state space*

- A is the *action space* (finite)

- $P : S \times A \to \Delta(S)$ is the *transition function* where the probability of reaching state s' from s, a is $\mathbb{P}(s' \mid s, a)$ for $\mathbb{P}$ defined by $P(s, a)$

- $r : S \times A \to [0, 1]$ is the immediate *reward function*

- $\gamma \in (0, 1)$ is the *discount factor*

- $\mu \in \Delta(S)$ is some *intial distribution*

that satisfies the *Markov property* or the *memoryless property*

$$\mathbb{P}(s' \mid \tau_t, a) = \mathbb{P}(s' \mid s_t, a)$$

Note that we are taking discounted *infinite horizon* MDPs, in which we interact with the environment infinitely many times.

Definition 3.3. The *history H* is the collection of all *trajectories $\tau_t = \{s_0, a_0, r_0, s_1, a_1, r_1, \cdots, s_t, a_t, r_t\}$.*

Definition 3.4. A *policy $\pi : H \to \Delta(A)$* is a decision making strategy in which the agent chooses actions adaptively. It is a *stationary policy* if $\pi : S \to \Delta(A)$ and a *deterministic policy* if $\pi : H \to A$.

Definition 3.5. The *value $V^\pi : S \to \mathbb{R}$* captures how valuable it is to be at a state s given policy π and is defined by

$$V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, \pi, s_0 = s\right]$$

Definition 3.6. The *Q-function $Q^\pi : S \times A \to \mathbb{R}$* captures the quality of a state s and action a given policy π and is defined by

$$Q^\pi(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, \pi, s_0 = s, a_0 = a\right]$$

Our goal for an agent in state s is to find a policy π that maximizes its value or $\pi^* \in \arg\max_{\pi \in \Pi} V^\pi(s)$.

Proposition 3.7 (Bellman equations for stationary policies).

$$Q^\pi(s, \pi(s)) = V^\pi(s)$$

$$Q^\pi(s, a) = r(s, a) + \gamma \mathop{\mathbb{E}}_{s' \sim \mathbb{P}(\cdot \mid s, a)} [V^\pi(s')]$$

Proof. The first equation is straightforward by considering that we get the action $\pi(s)$ from following π at state s. For the second equation,

$$Q^\pi(s, a) = \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t r(s_t, a_t) \,\middle|\, \pi, s_0 = s, a_0 = a\right] \tag{10}$$

$$= \mathbb{E}\left[r(s, a) + \sum_{t=1}^\infty \gamma^t r(s_t, a_t) \,\middle|\, \pi\right] \tag{11}$$

$$= r(s, a) + \gamma\mathbb{E}\left[\sum_{t=0}^\infty \gamma^t r(s_{t+1}, a_{t+1}) \,\middle|\, \pi\right] \tag{12}$$

$$= r(s, a) + \gamma\mathbb{E}\left[\underset{s'\sim\mathbb{P}(\cdot|s,a)}{\mathbb{E}}\left[\sum_{t=0}^\infty \gamma^t r(s_t, a_t) \,\middle|\, \pi, s_0 = s'\right]\right] \tag{13}$$

$$= r(s, a) + \gamma \underset{s'\sim\mathbb{P}(\cdot|s,a)}{\mathbb{E}}[V^\pi(s')] \tag{14}$$

where (13) comes from noting $s_1 \sim \mathbb{P}(\cdot \mid s, a)$ originally and shifting the indices by the Markov property. We then get to our result (14) by Fubini-Tonelli and the definition of $V^\pi(s')$. □

Theorem 3.8. *Let*

$$V^*(s) := \sup_{\pi\in\Pi} V^\pi(s)$$

$$Q^*(s, a) := \sup_{\pi\in\Pi} V^\pi(s, a).$$

Then, $\exists\tilde{\pi} \in \Pi_{stat,\ det}$ s.t. $\forall s, a, \in S \times A$,

$$V^{\tilde{\pi}}(s) = V^*(s)$$

$$Q^{\tilde{\pi}}(s, a) = Q^*(s, a).$$

Theorem 3.9 (Bellman optimality equation). *We say $Q : S \times A \rightarrow [0, 1]$ satisifes the Bellman optimality equation if $\forall s, a \in S \times A$,*

$$Q(s, a) = r(s, a) + \gamma \underset{s'\sim\mathbb{P}(\cdot|s,a)}{\mathbb{E}}\left[\max_{a'} Q(s', a')\right].$$

Then, $Q = Q^* \iff Q$ satisfies the Bellman optimality equation. Furthermore, policy $\pi, s \mapsto \arg\max_a Q^*(s, a)$ is an optimal policy.

Proof. We start by showing $V^*(s) = \max_a Q^*(s, a)$. Let $\pi^* \in \Pi_{stat,\ det}$ be an optimal policy as per previous theorem. □

Theorem 3.10 (Bellman Operator). *The Bellman Operator is $T^* : (S \times A \rightarrow [0, 1]) \rightarrow (S \times A \rightarrow [0, 1])$ s.t.*

$$(T^*f)(s, a) = r(s, a) + \gamma \underset{s'\sim\mathbb{P}(\cdot|s,a)}{\mathbb{E}}\left[\max_{a'} f(s', a')\right]$$

Remark 3.11. We have $T^*Q^* = Q^*$, which gives a way to find Q^* by repeatedly applying the Bellman operator. Note that T^* is also contractive in the sup norm, or $\|T^*f - T^*g\|_\infty \leq \gamma \|f - g\|_\infty$

Definition 3.12 (Greedy). The greedy policy π_Q is defined by $s \mapsto \arg\max_a Q(s, a)$.

Lemma 3.13 (Singh & Yee). $\forall Q : S \times A \to [0, 1]$,

$$V^{\pi_Q} \geq V^* - \frac{2\|Q - Q^*\|_\infty}{1 - \gamma}$$

Proof. For a state $s \in S$, we have

$$
\begin{align}
V^*(s) - V^{\pi_Q}(s) =& Q^*(s, \pi^*(s)) - Q^{\pi_Q}(s, \pi_Q(s)) \tag{15} \\
=& Q^*(s, \pi^*(s)) - Q^*(s, \pi_Q(s)) + Q^*(s, \pi_Q(s)) - Q^{\pi_Q}(s, \pi_Q(s)) \tag{16} \\
=& Q^*(s, \pi^*(s)) - Q^*(s, \pi_Q(s)) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s, \pi_Q(s))}[V^{\pi_Q}(s') - V^*(s')] \tag{17} \\
=& Q^*(s, \pi^*(s)) - Q(s, \pi^*(s)) + Q(s, \pi_Q(s)) - Q^*(s, \pi_Q(s)) + \tag{18} \\
& \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s, \pi_Q(s))}[V^{\pi_Q}(s') - V^*(s')] \\
\leq& 2\|Q - Q^*\|_\infty + \gamma\|V^* - V^{\pi_Q}\|_\infty \tag{19}
\end{align}
$$

where (3) comes from expanding by the Bellman equations and the linearlity of $\mathbb{E}$. (4) we have from $Q(s, \pi^*(s)) \leq Q(s, \pi_Q(s))$ by optimality, and (5) we get from the definition of the sup norm. Then, we have

$$(V^*(s) - V^{\pi_Q}(s)) - \gamma\|V^* - V^{\pi_Q}\|_\infty \leq 2\|Q - Q^*\|_\infty$$
$$\implies (1 - \gamma)(V^*(s) - V^{\pi_Q}(s)) \leq 2\|Q - Q^*\|_\infty$$
$$\implies V^*(s) - V^{\pi_Q}(s) \leq \frac{2\|Q - Q^*\|_\infty}{1 - \gamma}$$

by the definition of the sup norm again and moving terms around. $\square$

Value Iteration

Theorem 3.14. *Let* $Q^0 = 0, k \in \mathbb{N}$ *and suppose* $Q^{k+1} = T^*Q^k$. *Let* $\pi^k = \pi_{Q^k}$. *Then, for* $k \geq \frac{1}{1-\gamma}\log\frac{2}{(1-\gamma^2)\varepsilon}$,

$$V^{\pi^k} \geq V^* - \varepsilon$$

Proof. We know $\|Q^*\|_\infty \leq \frac{1}{1-\gamma}$, $Q^k = (T^*)^k Q^0$ and $Q^* = T^*Q^*$.

$$
\begin{align*}
\|Q^k - Q^*\|_\infty &= \|(T^*)^k Q^0 - (T^*)^k Q^*\|_\infty \\
&\leq \gamma^k \|Q^0 - Q^*\|_\infty \\
&\leq (1 - (1 - \gamma))^k \frac{1}{1 - \gamma} \\
&\leq \frac{\exp(-(1-\gamma)k)}{1 - \gamma}
\end{align*}
$$

$\square$

Policy Iteration

Theorem 3.15. *Let* π_0 *be any policy. For* $k \geq \frac{1}{1-\gamma}\log\frac{1}{(1-\gamma)\varepsilon}$, *the kth policy in Policy Iteration has*

$$V^{\pi^*} \geq V^* - \varepsilon$$

Prediction & Control

How can we control?

1. Model Based Methods: Estimating $\hat{P}, \hat{r}$

$$\hat{\mathbb{P}}(s' \mid s, a) \approx \frac{\text{count}(s, a, s')}{N(s, c)}$$

2. Model Free Methods: Estimating V^{π}
   Prediction: given π, how can we find V^{π}? Our goal is to find a V such that $V \approx V^{\pi}$. So, we will play policy π where we update

$$S_{t+1} = r_t + \gamma V_t(s_{t+1}) - V_t(s_t)$$

$$V_{t+1}(s) \leftarrow V_t(s) + \alpha_t S_{t+1}$$

over steps t. This is called *temporal difference (TD)* learning where S_{t+1} is the TD error. We want $\mathbb{E}[S_{t+1} \mid \cdots] = 0$. Define

$$\text{FV}(s) := \mathbb{E}[r_t + \gamma V_t(s_{t+1}) - V_t(s_t) \mid s_t = s]$$

Assume the Robbins-Monroe condition on $\{\alpha_t : t \geq 0\}$ i.e. $\sum \alpha_t = \infty, \sum \alpha_t^2 < \infty$. The sequence $\{V_t : t \geq 0\}$ traces out the trajectory of the ODE $\frac{\mathrm{d}}{\mathrm{d}t} v(t) = K \cdot \text{FV}(t)$ for constant K, which is asymptotically stable and converges a.s. to V^{π} (Szepesvari, 2009).