Supplementary Material for "Sample Efficient Reinforcement Learning with Gaussian Processes"

Robert C. Grande RGRANDE@MIT.EDU

Thomas J. Walsh THOMASJWALSH@GMAIL.COM

Jonathan P. How JHOW@MIT.EDU

Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139 USA

1. Full Proof of Lemma 1

Lemma 1 Consider a GP trained on samples $\vec{y} = [y_1, \ldots, y_t]$ which are drawn from $p(y \mid x)$ at input locations $X = [x_1, \ldots, x_t]$, with $\mathbb{E}[y \mid x] = f(x)$ and $V_m = y_{max} - y_{min}$. If the predictive variance of the GP at $x_i \in X$ is

$$\sigma^2(x_i) \leq \sigma^2_{tol} = \frac{2\omega_n^2 \epsilon_1^2}{V_m^2 \log(\frac{2}{\delta_1})} \tag{1}$$

then the prediction error at x_i is bounded in probability: $Pr\left\{|\mu(x_i) - f(x_i)| \geq \epsilon_1\right\} \leq \delta_1$.

Proof In order to prove that the GP estimate concentrates around the mean $f(x)$. We first use McDiarmid's Inequality to show that if the variance at a point satisfies 1, then the estimate of the GP is concentrated around its expected value with high probability. Secondly, since it is known that GPs are consistent estimators (Rasmussen & Williams, 2006), it follows that the expected value of the GP is the expected value of the distribution $f(x)$.

McDiarmid's Inequality states that

$$Pr\{|f(x_1, \ldots, x_n) - \mathbb{E}[f(x_1, \ldots, x_n)]| \leq \epsilon\} \leq 2\text{exp}\left(-\frac{2\epsilon^2}{\sum_i c_i^2}\right) = \delta \tag{2}$$

where $c_i = \sup f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, \hat{x}_i, \ldots, x_n)$. I.e. replacing x_i by some other value $\hat{x}_i$ can result in a change in the output $f(x_1, \ldots, x_n)$ no larger than c_i. In the case of an average of the variables, McDiarmid's Inequality becomes Hoeffding's Inequality. Consider the general GP regression equations

$$\mu(X) = K(X, X)(K(X, X) + \omega_n^2 I)^{-1} y \tag{3}$$

where $y \in [0, V_m]$ and $\text{Var}(y) \leq V_m^2$. $K(X, X)$ is symmetric and positive semi-definite, so its eigenvectors are orthonormal to eachother and all of its eigenvalues are nonnegative. It can be shown that $K(X, X)$ and $(K(X, X) + \omega_n^2)^{-1}$ have the same eigenvectors. Performing eigendecomposition,

$$\mu(X) = Q\Lambda Q^T Q(\Lambda + \omega_n^2 I)^{-1} Q^T \vec{y} \tag{4}$$

$$\mu(X) = Q\Lambda(\Lambda + \omega_n^2 I)^{-1} Q^T \vec{y} \tag{5}$$

Consider performing prediction only at the first input location x_1 by premultiplying using a unit coordinate vector $e_1 = [1, 0, \ldots, 0]^T$.

$$\mu(x_1) = e_1^T Q\Lambda(\Lambda + \omega_n^2 I)^{-1} Q^T \vec{y} \tag{6}$$

This is just a weighted sum of the observations y, with weights given by

$$\alpha = e_1^T Q\Lambda(\Lambda + \omega_n^2 I)^{-1} Q^T \tag{7}$$

It follows that $\sum_i c_i^2 = \|\alpha\|_2^2 V_m^2$. We have that

$$\|\alpha\|_2^2 = e_1^T Q\Lambda(\Lambda + \omega_n^2 I)^{-1} Q^T Q(\Lambda + \omega_n^2 I)^{-1}\Lambda Q^T e_1 \tag{8}$$

$$\|\alpha\|_2^2 = q_1\Lambda(\Lambda + \omega_n^2 I)^{-1}(\Lambda + \omega_n^2 I)^{-1}\Lambda q_1^T \tag{9}$$

where $q_1 = [Q_{11} \dots Q_{1n}]$ is the first row of Q. Therefore, we have

$$\|\alpha\|_2^2 = \sum_i q_{1i}^2 \left(\frac{\lambda_i}{\lambda_i + \omega^2}\right)^2 \tag{10}$$

However, by evaluating (7), we have that the weight α_1 which corresponds to (x_1, y_1) is given by

$$\alpha_1 = \sum_i q_{1i}^2 \frac{\lambda_i}{\lambda_i + \omega^2}. \tag{11}$$

Since every term in (11) is greater than every respective term in the sum of (10), it follows that,

$$\|\alpha\|_2^2 \leq \alpha_1 \tag{12}$$

In order to finish the proof, we now upper bound α_1. Consider that a GP prediction is equivalent to the MAP estimate of a linear gaussian measurement model with a gaussian prior.

$$\mu_{MAP}(x_1) = \frac{\sigma_0^2(x_1)}{\omega^2 + \sigma_0^2(x_1)} y_1 + \frac{\omega^2}{\sigma_0^2(x_1) + \omega^2}\mu_0(x_1) \tag{13}$$

In this case, the prior mean $\mu_0(x_1)$, and variance $\sigma_0^2(x_1)$ are given by the GP estimate before including (x_1, y_1), and the weight of the new observation is given by

$$\alpha_1 = \frac{\sigma_0^2(x_1)}{\omega^2 + \sigma_0^2(x_1)} \leq \frac{\sigma^2(x_1)}{\omega_n^2} \tag{14}$$

Using this bound on α_1, we have from McDairmid's Inequality that if

$$\frac{1}{\sigma^2(x_1)} = \frac{V_m^2}{2\omega_n^2\epsilon_1^2}\log(\frac{2}{\delta}) \tag{15}$$

then our prediction is within ϵ_1 of the expected value of the GP prediction with probability $1-\delta$. This proves that the estimate of the GP concentrates around its expected value with high probability. Since GPs are consistent estimators (Rasmussen & Williams, 2006), it follows that the expected value of the GP is the expected value of the distribution $f(x)$. Therefore, it follows that if (15) holds, then the estimate of the GP is within ϵ_1 of $f(x)$.

1.1. Role of the Prior in Lemma 1

This section describes the affect of the GP prior on the KWIK learning result in Lemma 1.

Traditionally, the KWIK framework does not incorporate prior model information in a Bayesian manner, which contrasts with the use of priors by a GP. While priors may improve early model performance, they may also slow learning if the initial estimate is far from the actual value. In order to consider the effect of the prior on GP inference, we examine the bias versus variance properties of the GP estimator for properties of the model. The mean function of a GP is equivalent to the MAP estimate of a linear combination of Gaussian beliefs, given by,

$$\hat{\mu} = \frac{\sigma_0^2}{\sigma^2 + \sigma_0^2} f(y_1, \dots, y_n) + \frac{\sigma^2}{\sigma^2 + \sigma_0^2}\mu_0 \tag{16}$$

where μ_0, σ_0^2 are the prior mean and variance, respectively. $f(\cdot)$ is a function that maps the observations to an estimate. In the single variable case, this corresponds to an average. In Lemma 1, this corresponds to the function $\mu(x') = f(x_1, \dots, x_n)$, which was analyzed. Therefore, the total error induced by using a GP is given

by the sensitivity (variance) of $f(\cdot)$ to noise, which was analyzed in Lemma 1, and the error induced by having a prior bias. Here, we show that this bias-error is small, and can be effectively ignored in the analysis.

The maximum error due to regularization is given by the upper bound on the second term

$$\epsilon_{pr} = \frac{\sigma^2}{\sigma^2 + \sigma_0^2} V_m \tag{17}$$

Plugging in σ_{tol} from Lemma 1,

$$\epsilon_{pr} = \frac{\omega_n^2 \epsilon_1^2 V_m}{\omega_n^2 \epsilon_1^2 + \frac{1}{2} V_m^2 \log(\frac{2}{\delta_1}) \sigma_0^2} \tag{18}$$

Defining $r = \frac{\sigma_0^2}{\omega_n^2}$,

$$\epsilon_{pr} = \frac{V_m}{1 + \frac{r}{2\epsilon_1^2} V_m^2 \log(\frac{2}{\delta_1})} \tag{19}$$

Interestingly, the role of the prior decreases as V_m, $\frac{1}{\epsilon_1}$, and $\frac{1}{\delta}$ increase. This is due to the fact that the variance required to average out noise scales as V_m^2 whereas the error induced by bias of the prior scales V_m. The effect of the prior increases as the ratio r decreases, i.e. our initial variance is very low, or the measurement noise is high. We argue that by increasing r, we can make the error due to regularization arbitrarily small without effecting the reliability of our estimate. For example, by increasing σ_0^2 arbitrarily large, we have,

$$\hat{\mu} \approx \frac{\sigma_0^2}{\sigma^2 + \sigma_0^2} f(y_1, \ldots, y_n) \tag{20}$$

Effectively, by making the prior variance arbitrarily large, we have turned the GP into a function approximator that does not rely on any prior information. Therefore the guarantees still hold from the previous section. By creating a large prior variance, one might be concerned that the rate at which σ^2 decays will be slowed down. However, consider that the covariance at a point after a single observation is given by $\sigma^2 = (\sigma_0^{-2} + \omega^{-2})^{-1} \approx \omega^2$. Thus, for any $\sigma_0^2 >> \omega^2$, the effect of the prior variance does not matter much. Since we can set r arbitrarily large by scaling the kernel function, we can therefore scale ϵ_{pr} to be arbitrarily small. Therefore, we neglect the role of the prior in the analysis.

2. Theorem 1: Proof of the variance decrease rate in a Voronoi region

An ϵ_t-volume around a point $\bar{x}$ is defined as the set of all points which satisfy the following distance metric as $\{\epsilon_t - \text{Vol} : x \in S \mid k(\bar{x}, \bar{x}) - k(\bar{x}, x)^T K(x, x)^{-1} k(\bar{x}, x)\}$. Define the correlation coefficients between two points as $\rho = k(x_i, x_j)$. Using bayes law, it can be shown that given a point $\bar{x}$ with prior uncertainty $\sigma_0^2 = 1$ and m measurements at another location x_i with correlation coefficient ρ, the posterior variance is given by $\sigma_n^2 = 1 - \frac{n\rho^2}{n+\omega^2}$. Using the linear independence test, we relate ϵ_t to ρ as $\epsilon_t = 1 - \rho^2$. Therefore, we have that in an ϵ_t-volume, the slowest that the variance can reduce at the center of the volume $\bar{z}$ is given by,

$$\sigma_n^2 \leq \frac{n\epsilon_t + \omega^2}{n + \omega^2} \leq \frac{n\epsilon_t + \omega^2}{n} \tag{21}$$

The ϵ_t-volume around a point $\bar{x}$ is identical to the voronoi region around a point in the covering set. $\epsilon_t \leq \frac{1}{2}\sigma_{tol}^2$ by the definition of the covering number.

3. Theorem 1: Proof that $\mathcal{N}_U(r)$ grows polynomially

Theorem 1 Furthermore, $\mathcal{N}_U(r(\epsilon_{tol}))$ grows polynomially with $\frac{1}{\epsilon_1}$ and V_m for the Radial Basis Function (RBF) kernel.

Proof We can bound the covering number $\mathcal{N}_U(r(\frac{1}{2}\sigma_{tol}^2))$ loosely by creating a hyper-parallelopiped, which countains the entire state space X, with dimensions $l_1, l_2, \ldots, l_d$, where d is the dimension of the space. The covering

number is then loosely bounded by dividing the volume of the hyper-parallelopiped by hyper-cubes of dimension $r(\frac{1}{2}\sigma_{tol}^2)$, which is a strictly smaller than the true volume of each voronoi region. Plugging in,

$$\mathcal{N}_U(r(\frac{1}{2}\sigma_{tol}^2)) = \frac{l_1 l_2 \dots l_d}{r(\frac{1}{2}\sigma_{tol}^2)^d} \tag{22}$$

In the case of the RBF kernel $k(x, x') = \exp(-\frac{\|x-x'\|^2}{2\theta})$, the equivalent distance map is given by

$$r(\epsilon_{tol}) = \theta \left(\log \left(\frac{1}{1-\epsilon_{tol}} \right) \right)^{\frac{1}{2}}. \tag{23}$$

$\frac{1}{r(\epsilon_{tol})}$ grows polynomially with $\frac{1}{\epsilon_{tol}}$. Additionally, $\frac{1}{\epsilon_{tol}} = \frac{1}{2}\sigma_{tol}^2 = \frac{V_m^2 \log\left(\frac{2}{\delta}\right)}{\omega^2 \epsilon_1^2}$, so it follows, $\mathcal{N}_U(r(\frac{1}{2}\sigma_{tol}^2)) \sim O\left(f^p(V_m^2, \frac{1}{\epsilon_1^2}, \log\left(\frac{1}{\delta}\right))\right)$, where $f^p(\cdot)$ is some function bounded by a polynomial of order p.

4. Proof of Lemma 2

Lemma 2 The total number of successful updates (overwrites) during any execution of DGPQ is

$$\kappa = |A|\mathcal{N}_S\left(\frac{\epsilon(1-\gamma)}{3L_Q}\right)\left(\frac{3R_{min}}{(1-\gamma)^2\epsilon} + 1\right) \tag{24}$$

Proof We begin with the single state/action example from Section 5 and then expand on this result. There, given an estimate $\hat{Q}_i$, the corresponding difference in measurements is $(1-\gamma)\hat{Q}_i$. The stopping criterion for the algorithm is given by

$$(1-\gamma)\hat{Q}_i \leq 2\epsilon_1 \tag{25}$$

The error associated with stopping the algorithm at this stage as opposed to running the algorithm to infinity is given by the infinite sum of the geometric series,

$$\epsilon = \epsilon_1 + \sum_{i=0}^{\infty}(2\epsilon_1)\gamma^i = \epsilon_1 + \frac{2\epsilon_1}{1-\gamma} \tag{26}$$

The maximum number of updates, η, required to reach this stopping point is given by the maximum distance the estimate $\hat{Q}$ can move, 1, divided by the minimum swap distance ϵ_1

$$\eta = \frac{1}{\epsilon_1} \tag{27}$$

$$\eta = \frac{3-\gamma}{\epsilon(1-\gamma)} \tag{28}$$

$$\eta \leq \frac{3}{\epsilon(1-\gamma)} \tag{29}$$

Similarly, in a multi-state domain, the number of overwrites required to bring one point down from $\hat{Q}(s,a) = \frac{R_{max}}{1-\gamma}$ to $\hat{Q}(s,a) = \epsilon$ is given by $\eta = \frac{3R_{max}}{(1-\gamma)^2\epsilon}$. Manipulating (26), $\epsilon_1 = \frac{1}{3}\epsilon(1-\gamma)$ satisfies a final error of ϵ at (s,a). Consider the worst case analysis in which $Q^*(s,a) = \frac{R_{max}}{1-\gamma}$, $\forall (s,a) \in U$. Due to the representation of $\hat{Q}(s,a)$, there is also an error induced by optimism away from the basis vector location. This error is given by $\epsilon_\Delta = L_Q d((s,a),(s_i,a_i))$. If the furthest point in the basis vector set is at most $r = \frac{\epsilon(1-\gamma)}{3L_Q}$ (given by the definition of the covering set), then $\epsilon_\Delta = \epsilon_1$. The worst case analysis is then that the algorithm updates $\hat{Q}$, η times multiplied by the covering number $\mathcal{N}_c(\frac{\epsilon(1-\gamma)}{3L_Q})$ in order to bring $|\hat{Q}(s,a) - Q^*(s,a)| \leq \epsilon + \epsilon_1$ everywhere. The algorithm then can update $\hat{Q}$ at most once more per covering volume before $|\hat{Q}(s,a) - Q^*(s,a)| \leq \epsilon$ everywhere, leading to a final number of $\hat{Q}$ updates of $\kappa = |A|\mathcal{N}_S(\frac{\epsilon(1-\gamma)}{3L_Q})\left(\frac{3R_{max}}{(1-\gamma)^2\epsilon} + 1\right)$

5. Proof of Lemma 3

Lemma 3 The total number of **attempted** updates (overwrites) during any execution of GPQ is $|A|\mathcal{N}_S\left(\frac{\epsilon(1-\gamma)}{3L_Q}\right)(1+\kappa)$.

Proof After an overwrite has occurred and the GPs are reinitialized, the variance can fall below σ^2_{tol} a maximum of $|A|\mathcal{N}_S\left(\frac{\epsilon(1-\gamma)}{3L_Q}\right)$ times. There are a maximum of κ successful updates. After κ updates, the variance can fall below σ^2_{tol} a maximum of $|A|\mathcal{N}_S\left(\frac{\epsilon(1-\gamma)}{3L_Q}\right)$ times before the variance has fallen below σ^2_{tol} everywhere.

6. Proof of Lemma 5

Lemma 4 During execution of DGPQ, $Q^*(s,a) \leq Q_t(s,a) + \frac{2\epsilon_1}{1-\gamma}$ holds for all $\langle s,a\rangle$ with probability $\frac{\delta}{3}$.

Proof Define $BQ(s,a) = \mathbb{E}[r(s,a) + \gamma \max_{a'} Q(s',a')]$ to be the exact Bellman update of a value function, and $\tilde{B}Q(s,a)$ to be the approximate Bellman operator using the update of $\hat{Q}$ from Algorithm 1. From Lemma 1, setting σ^2_{tol} as per (1), and performing an overwrite as per Algorithm 1, we have that for every update, $BQ(s,a) \leq \tilde{B}Q(s,a)$ w.p. $1 - \frac{1}{3|A|\mathcal{N}_S\left(\frac{\epsilon(1-\gamma)}{3L_Q}\right)(1+\kappa)}\delta$. This is due to the fact that $|GP_a.\text{mean}(s_t) - BQ(s,a)| \leq \epsilon_1$ w.p. $1 - \frac{1}{3|A|\mathcal{N}_S\left(\frac{\epsilon(1-\gamma)}{3L_Q}\right)(1+\kappa)}\delta$, and that we add a bonus term of ϵ_1 when we perform an update.

From Lemma 3.10 of (Pazis & Parr, 2013), if $\hat{Q}(s,a)$ is initialized optimistically, and the Bellman operator is only applied if the difference between the Bellman update and the current value of the Q-function differs by more than $2\epsilon_1$, i.e. $BQ_t(s,a) \leq Q_t(s,a) + 2\epsilon_1$, then we have that our estimate of the Q-function is always optimistic as $Q^*(s,a) \leq Q_t(s,a) + \frac{2\epsilon_1}{1-\gamma}$, $\forall t$.

If $BQ(s,a) \leq \tilde{B}Q(s,a)$ holds for all updates, then it follows that our approximation is optimistic at the update locations: $Q^*(s,a) \leq \hat{Q}_t(s,a) + \frac{2\epsilon_1}{1-\gamma}$. For any other location (s,a), Equation 9 adds a bonus of $L_Q \times d$, where L_Q is the lipschitz constant, and d is the distance. Since Q^* cannot grow faster than the lipschitz constant, this bonus ensures that the estimate of Q^* remains optimistic away from known locations. The probability that DGPQ does not remain optimistic can be conservatively estimated as the union bound of the probability of any update failing, which is given by $\delta/3$.

7. Proof of Lemma 6

Lemma 5 If event A2 occurs, then if an unsuccessful update occurs at time t and $GP_a.\text{Var}(s) < \sigma^2_{tol}$ at time $t+1$ then $\langle s,a\rangle \in K_{t+1}$.

Proof Suppose the conditions in the lemma hold but $\langle s,a\rangle \notin K_{t+1}$. Since the update was unsuccessful we know that $\langle s,a\rangle \notin K_t$ as well. However, because A2 occurred and the update was unsuccessful, we know there was some $k_1 < t$ where $\langle s,a\rangle \in K_{k_1}$. But this would mean that a successful update occurred between k_1 and t, which means GP_a would have been reset, so $GP_a.\text{Var}(s) \geq \sigma^2_{tol}$, which is a contradiction.

8. Proof of Lemma 7

Lemma 6 If event A2 occurs and $\hat{Q}_t(s,a) \geq Q^*(s,a) - \frac{2\epsilon_1}{1-\gamma}$ holds for all t and $\langle s,a\rangle$ then the number of timesteps ζ where $\langle s_t,a_t\rangle \notin K_t$ is at most

$$\zeta = m|A|\mathcal{N}_S\left(\frac{\epsilon(1-\gamma)}{3L_Q}\right)\left(\frac{3R_{max}}{(1-\gamma)^2\epsilon}+1\right) \tag{30}$$

where

$$m = \left(\frac{36R^2_{min}}{(1-\gamma)^4\epsilon^2}\log\left(\frac{6}{\delta}|A|\mathcal{N}_S\left(\frac{\epsilon(1-\gamma)}{3L_Q}\right)(1+\kappa)\right)\right)|A|\mathcal{N}_S\left(\frac{\epsilon(1-\gamma)}{3L_Q}\right) \tag{31}$$

Proof The number of timesteps ζ can be decomposed into two parts, the maximum number of successful updates $\kappa = |A|\mathcal{N}_S\left(\frac{\epsilon(1-\gamma)}{3L_Q}\right)\left(\frac{3R_{max}}{(1-\gamma)^2\epsilon}+1\right)$, and the maximum number of times (after an update has occurred) that the agent will encounted a state with high variance, i.e. the state is unknown, $m = \left(\frac{36R_{min}^2}{(1-\gamma)^4\epsilon^2}\log\left(\frac{6}{\delta}|A|\mathcal{N}_S\left(\frac{\epsilon(1-\gamma)}{3L_Q}\right)(1+\kappa)\right)\right)|A|\mathcal{N}_S\left(\frac{\epsilon(1-\gamma)}{3L_Q}\right)$. κ is taken from Lemma 2. m is obtained by plugging in $\delta_1 = \frac{\delta}{3|A|\mathcal{N}_S\left(\frac{\epsilon(1-\gamma)}{3L_Q}\right)(1+\kappa)}$, $\epsilon_1 = \frac{1}{2}\epsilon(1-\gamma)$ into Equation 3 of Theorem 1.

There are two cases for $\langle s_t, a_t\rangle \notin K_t$, which by Lemma 5 means that $GP_{a_t}.Var(s_t) > \sigma_{tol}^2$ before the GP update.

First, if $GP_{a_t}.Var(s) < \sigma_{tol}^2$ after the GP update, then an attempted update to $\hat{Q}$ will occur at time t, and must be successful because the state has an erroneous value (since it is not in K_t) and event A2 has occurred. Therefore, this case can only happen κ times by Lemma 2.

Second, if the variance $GP_{a_t}.\mathrm{Var}(s_t) > \sigma_{tol}^2$ after the GP update, then the next attempted update must occur within m encounters of unknown states, since by then whichever state is $\notin K_t$ must reach convergence in the GP. This is because each encounter of an unknown state/action when event A2 occurs is also an encounter of a state/action with $GP_{a_t}.\mathrm{Var}(s_t) > \sigma_{tol}^2$.

References

Pazis, Jason and Parr, Ronald. PAC optimal exploration in continuous space markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2013.

Rasmussen, C. and Williams, C. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.