数理解析研究所講究録
第2078巻 2018年 102-106

Bayes-optimal Fine-tuning under Imperfect Observations

Jue Wang
Assistant Professor, Smith School of Business,
Queen's University, Kingston, ON, Canada

1 Introduction

Fine-tuning is a sequential adjustment process that brings a system to the highest level of performance. Automatic fine-tuning capability is crucial for smart products, which can learn from user's behavior to improve product performance and increase customer satisfaction. Suppose the output is a concave function of the input and can be represented by a quadratic response function. The output Y_t is observed with a normal noise, i.e.,

$$Y_t = -\alpha x_t^2 + \beta x_t + \epsilon_t, \tag{1}$$

where $x_t \in \mathbb{R}$ is the input variable chosen by the decision maker at time t, and $Y_t \in \mathbb{R}$ is the random output observed at the same time. Note that $\alpha > 0$ is required to ensure the concavity of the response function. The noise ϵ_t is iid. following the normal distribution $N(0, \tau)$, in which $\tau > 0$ is the precision. For tractability, we assume that α and τ are both known. Only β is unknown.

The unknown parameter, β, can be learned online in a Bayesian fashion. That is, we assume a prior belief $\pi_0(\beta)$ at the beginning of the decision horizon ($t = 0$), and update the posterior belief using Bayes' rule as more observations of x_t and Y_t become available.

Let Δ denote the set of all admissible policies. The decision maker employing the policy $\delta \in \Delta$ chooses the next action according to $x_{t+1} = \delta(I_t)$, where $I_t = \{\pi_0, x_1, y_1, \ldots, x_t, y_t\}$ is all the information obtained up to the time of decision. The decision maker's objective is to find a control policy that maximizes the expected total output over a finite horizon given the prior, namely,

$$\max_{\delta \in \Delta} \mathbb{E}_\delta \Big[\sum_{t=1}^{T} Y_t | \pi_0(\beta) \Big]. \tag{2}$$

When β is known, this problem is trivial because the optimal value of x_t is $\beta/(2\alpha)$, and the maximum expected output is $T\beta^2/(4\alpha)$. However, when β is unknown, the decision maker may explore its true value by varying the inputs x_t and observe the response. But too much exploration may forgo the opportunity to generate more outputs. Such tension between exploration and exploitation is the focus of this paper. Our model is a stylized one, with only one unknown parameter, and our focus is the characterization of the optimal exploration-and-exploitation policy.

Our model is related to the sequential sampling problem formulated by Bertsekas (1976), cf. Chapter 4.6 of [1]. However, Bertsekas (1976) does not study the structure of the optimal control policy. Our problem may appear similar to the multi-armed bandit (MAB) problem [2] in the sense that we are searching for the best input (or arm) which generates the highest output. However, it is different from the classic MAB problem because sampling each input also yields information about other input, as the parameter of each arm is related to the parameters of other arms in a quadratic form.

For mathematical tractability, we make use of the natural conjugate prior for normal distribution, which is also a normal density,

$$p(\beta | \mu_0, \tau_0) = \sqrt{\frac{\tau_0}{2\pi}} \exp\Big\{ -\frac{\tau_0(\beta - \mu_0)^2}{2} \Big\},$$

where the hyperparameters μ_0, τ_0 represent the mean and precision of the normal prior, respectively. Suppose the decision maker has observed a sequence of input and output data $\{x_1, y_1, \ldots, x_n, y_n\}$ up

to period n. The likelihood function is given by

$$p(y_1,\ldots,y_n|x_1,\ldots,x_n,\beta)=\Big(\frac{\tau}{2\pi}\Big)^{\frac{n}{2}}\exp\Big\{-\frac{\tau}{2}\sum_{t=1}^{n}\big(y_t+\alpha x_t^2-\beta x_t\big)^2\Big\},$$

The posterior is given by

$$\begin{aligned}&p(\beta|y_1,\ldots,y_n,x_1,\ldots,x_n,\pi_0)\propto p(y_1,\ldots,y_n|x_1,\ldots,x_n,\beta)p(\beta|\mu_0,\tau_0)\\&\propto\exp\Big\{-\frac{1}{2}\beta^2\big(\tau_0+\sum_{t=1}^{n}\tau x_t^2\big)+\beta[\tau_0\mu_0+\sum_{t=1}^{n}\tau x_t(y_t+\alpha x_t^2)]\Big\}.\end{aligned}\tag{3}$$

It is worth mentioning that the posterior is a function of β proportional to the terms listed in (3) through a normalization constant. The posterior is also a normal distribution with the following hyperparameters

$$\begin{aligned}\tau_n&=\tau_0+\sum_{t=1}^{n}\tau x_t^2,\\\mu_n&=\frac{\tau_0\mu_0+\sum_{t=1}^{n}\tau x_t(y_t+\alpha x_t^2)}{\tau_n}.\end{aligned}$$

2 Bayesian Dynamic Programming Formulation

Let $V_n(\tau_n,\mu_n)$ denote the maximum expected output-to-go at the period $n=1,\ldots,T$, given the posterior hyperparameters (τ_n,μ_n). At period n, the decision maker chooses input for the next period x_{n+1}. The optimality equations are given by

$$\begin{aligned}&V_n(\tau_n,\mu_n)=\max_{x_{n+1}}\big\{\mathbb{E}[Y_{n+1}|x_{n+1},\mu_n]+\bar{V}_{n+1}(x_{n+1},\tau_n,\mu_n)\big\},\\&V_T(\tau_T,\mu_T)=0,\end{aligned}\tag{4}$$

where $\mathbb{E}[Y_{n+1}|x_{n+1},\mu_n]=-\alpha x_{n+1}^2+\mu_n x_{n+1}$ is the expected output in period $n+1$ given the information available at period n and the chosen input, x_{n+1}. It only depends on the posterior mean, μ_n, not the precision, τ_n. $\bar{V}_{n+1}(x_{n+1},\tau_n,\mu_n)$ is the expected output-to-go after period $n+1$ given the input x_{n+1}. More specifically,

$$\begin{aligned}&\bar{V}_{n+1}(x_{n+1},\tau_n,\mu_n)\\&=\int_{-\infty}^{\infty}V_{n+1}\Big(\tau_n+\tau x_{n+1}^2,\frac{\tau_n\mu_n+\tau x_{n+1}(y_{n+1}+\alpha x_{n+1}^2)}{\tau_n+\tau x_{n+1}^2}\Big)f(y_{n+1}|x_{n+1},\tau_n,\mu_n)dy_{n+1},\end{aligned}\tag{5}$$

in which $f(y_{n+1}|x_{n+1},\tau_n,\mu_n)$ is the predictive density of the output at period $n+1$. It can be derived as the following:

$$f(y_{n+1}|x_{n+1},\tau_n,\mu_n)=\int_{-\infty}^{\infty}p(y_{n+1}|\beta,x_{n+1})p(\beta|\tau_n,\mu_n)d\beta=\frac{\exp\Big\{-\frac{(y_{n+1}+\alpha x_{n+1}^2-\mu_n x_{n+1})^2}{2(1/\tau+x_{n+1}^2/\tau_n)}\Big\}}{\sqrt{2\pi(1/\tau+x_{n+1}^2/\tau_n)}}.$$

This is a normal density with mean $-\alpha x_{n+1}^2+\mu_n x_{n+1}$ and variance $1/\tau+x_{n+1}^2/\tau_n$. It turns out that an alternative representation of the optimality equation is much easier to analyze. Instead of using y_{n+1} as the variable of integration in (5), we can make the following change of variable

$$\mu_{n+1}\triangleq\frac{\tau_n\mu_n+\tau x_{n+1}(y_{n+1}+\alpha x_{n+1}^2)}{\tau_n+\tau x_{n+1}^2},$$

and use the new variable μ_{n+1} as the variable of integration. In this way, (5) can be rewritten as

$$\bar{V}_{n+1}(x_{n+1},\tau_n,\mu_n)=\int_{-\infty}^{\infty}V_{n+1}\big(\tau_n+\tau x_{n+1}^2,\mu_{n+1}\big)\frac{\exp\Big\{-\frac{(\mu_{n+1}-\mu_n)^2}{2v_{n+1}}\Big\}}{\sqrt{2\pi v_{n+1}}}d\mu_{n+1},\tag{6}$$

where

$$v_{n+1} = \frac{1}{\tau_n} - \frac{1}{\tau_n + \tau x_{n+1}^2},$$

is the conditional variance of μ_{n+1} given the information available at period n. Here τ_n^{-1} is the variance of the belief before making a new observation, and $(\tau_n + \tau x_{n+1}^2)^{-1}$ is its variance after the new observation, obtained at the input x_{n+1}. Their difference, v_{n+1}, can be interpreted as the *reduction of uncertainty* by the new information. It is important to observe that the mean of μ_{n+1} is still μ_n. But the variance of μ_{n+1} is increasing in x_{n+1}^2 and decreasing in the posterior precision, τ_n.

Clearly, if $\tau = 0$ or $x_{n+1} = 0$, there will be no reduction in the variance of belief. In this case, the posterior mean is always identical to the prior mean and no learning will occur. If $\tau \to \infty$, the true parameter will be revealed at period $n+1$, the posterior mean will be equal to the true parameter. Note that we do not know the true parameter, so the updated posterior mean appears to be random at period n.

3 Optimal Fine-tuning Policy

In this section, we characterize the structure of the optimal policy. Some proofs are omitted for conciseness but are available from the author upon request. First, we show that $V_n(\tau_n, \mu_n)$ is convex and symmetric.

Proposition 1 (Element-wise convexity) *For any fixed τ_n, the value function $V_n(\tau_n, \mu_n)$ is convex in μ_n.*

Proposition 2 (Symmetry) *For any given τ_n, we have $V_n(\tau_n, \mu_n) = V_n(\tau_n, -\mu_n)$ for all μ_n. Further, $V_n(\tau_n, \mu_n)$ is increasing in μ_n when $\mu_n > 0$, decreasing when $\mu_n < 0$, and $V_n(\tau_n, \mu_n) \geqslant V_n(\tau_n, 0)$.*

The next result concerns the monotonicity of the value function in the precision of the belief. We use decreasing (increasing) and non-increasing (non-decreasing) interchangeably.

Proposition 3 (Monotonicity) *$V_n(\tau_n, \mu_n)$ is decreasing in τ_n for all n.*

The following lemma suggests that the value of information is positive and is decreasing in time.

Lemma 1 (Value of Information)

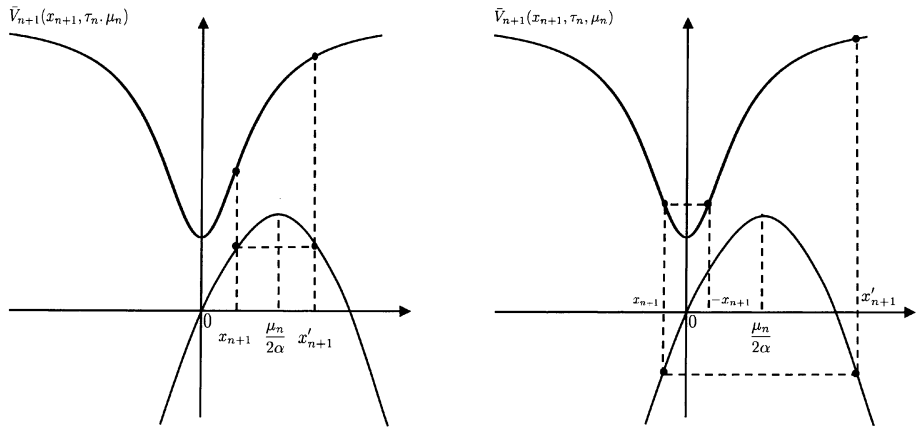
1. *$\bar{V}_{n+1}(x_{n+1}, \tau_n, \mu_n) \geqslant V_{n+1}(\tau_n, \mu_n)$ for all n, x_{n+1}, τ_n and μ_n.*
2. *$\bar{V}_{n+1}(x, \tau_n, \mu_n) - V_{n+1}(\tau_n, \mu_n) \geqslant \bar{V}_{n+2}(x, \tau_n, \mu_n) - V_{n+2}(\tau_n, \mu_n)$.*

Proposition 4 *For all τ_n, μ_n and n, the value function $\bar{V}_{n+1}(x_{n+1}, \tau_n, \mu_n)$ is increasing in x_{n+1} when $x_{n+1} > 0$, and decreasing in x_{n+1} when $x_{n+1} < 0$. Further, $\bar{V}_{n+1}(x_{n+1}, \tau_n, \mu_n) = \bar{V}_{n+1}(-x_{n+1}, \tau_n, \mu_n)$.*

Proof: Consider any two inputs x_a and x_b such that $x_a^2 > x_b^2 > 0$. To prove the piece-wise monotonicity, it suffices to prove that $\bar{V}_{n+1}(x_a, \tau_n, \mu_n) \geqslant \bar{V}_{n+1}(x_b, \tau_n, \mu_n)$ for all τ_n, μ_n, and n. A key step toward this goal is to express the expectation in $\bar{V}_{n+1}(x_a, \tau_n, \mu_n)$ by conditioning on a dummy variable μ'_{n+1}, as shown in the following

$$\bar{V}_{n+1}(x_a, \tau_n, \mu_n) \triangleq \int_{-\infty}^{\infty} V_{n+1}\big(\tau_n + \tau x_a^2, \mu_{n+1}\big) \frac{\exp\left\{ -\frac{(\mu_{n+1}-\mu_n)^2}{2v_{n+1}^a} \right\}}{\sqrt{2\pi v_{n+1}^a}} d\mu_{n+1},$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} V_{n+1}\big(\tau_n + \tau x_b^2 + \tau(x_a^2 - x_b^2), \mu_{n+1}\big) \frac{\exp\left\{ -\frac{(\mu_{n+1}-\mu'_{n+1})^2}{2(v_{n+1}^a - v_{n+1}^b)} \right\}}{\sqrt{2\pi(v_{n+1}^a - v_{n+1}^b)}} \frac{\exp\left\{ -\frac{(\mu'_{n+1}-\mu_n)^2}{2v_{n+1}^b} \right\}}{\sqrt{2\pi v_{n+1}^b}} d\mu_{n+1} d\mu'_{n+1},$$



Case I: $x_{n+1} \geqslant 0$ Case II: $x_{n+1} < 0$

Figure 1: Illustration of the Proof of Theorem 1.

where $v^a_{n+1} = 1/\tau_n - 1/(\tau_n + \tau x_a^2)$, $v^b_{n+1} = 1/\tau_n - 1/(\tau_n + \tau x_b^2)$. Next, note that Lemma 1 implies

$$\int_{-\infty}^{\infty} V_{n+1}\big(\tau_n + \tau x_b^2 + \tau(x_a^2 - x_b^2), \mu_{n+1}\big) \frac{\exp\left\{-\frac{(\mu_{n+1}-\mu'_{n+1})^2}{2(v^a_{n+1}-v^b_{n+1})}\right\}}{\sqrt{2\pi(v^a_{n+1}-v^b_{n+1})}} d\mu_{n+1} \geqslant V_{n+1}(\tau_n + \tau x_b^2, \mu'_{n+1}).$$

Therefore, we have

$$\bar{V}_{n+1}(x_a,\tau_n,\mu_n) \geqslant \int_{-\infty}^{\infty} V_{n+1}\big(\tau_n + \tau x_b^2, \mu_b\big) \frac{\exp\left\{-\frac{(\mu_b-\mu_n)^2}{2v^b_{n+1}}\right\}}{\sqrt{2\pi v^b_{n+1}}} d\mu_b = \bar{V}_{n+1}(x_b,\tau_n,\mu_n),$$

proving the piece-wise monotonicity. Finally, $\bar{V}_{n+1}(x_{n+1},\tau_n,\mu_n) = \bar{V}_{n+1}(-x_{n+1},\tau_n,\mu_n)$ follows immediately from (6). *QED.*

Remark 1 *A natural way to prove the monotonicity is by induction. However, this approach turns out to be difficult. When $n = T-1$, we have $\bar{V}_T(x_T,\tau_{T-1},\mu_{T-1}) = 0$, in which the monotonicity holds trivially. When $n = T-2$, it is easy to see from the expression that $\bar{V}_{T-1}(x_{T-1},\tau_{T-2},\mu_{T-2})$ is increasing in τx^2_{T-1}. To carry out the induction, we can suppose $\bar{V}_{n+1}(x_{n+1},\tau_n,\mu_n)$ is increasing in x^2_{n+1} and try to prove that $\bar{V}_n(x_n,\tau_{n-1},\mu_{n-1})$ is also increasing in x_n^2. Unfortunately, this argument does not work here. To see this, note that*

$$\bar{V}_{n+1}(x_{n+1},\tau_n,\mu_n) = \int_{-\infty}^{\infty} V_{n+1}\big(\tau_n + \tau x^2_{n+1}, \mu_{n+1}\big) \frac{\exp\left\{-\frac{(\mu_{n+1}-\mu_n)^2}{2v_{n+1}}\right\}}{\sqrt{2\pi v_{n+1}}} d\mu_{n+1}, \tag{7}$$

Although v_{n+1} is increasing in x^2_{n+1} and $V_{n+1}(\tau_n + \tau x^2_{n+1}, \mu_{n+1})$ is convex in μ_{n+1}, we cannot use the second-order stochastic dominance to prove that $\bar{V}_{n+1}(x_{n+1},\tau_n,\mu_n)$ is increasing in x^2_{n+1} here. This is because $V_{n+1}\big(\tau_n + \tau x^2_{n+1}, \mu_{n+1}\big)$ is decreasing *in x^2_{n+1} by Proposition 3.*

Define the optimal input as

$$x^\star_{n+1}(\mu_n,\tau_n) \triangleq \sup \operatorname{argmax}\{-\alpha x^2_{n+1} + \mu_n x_{n+1} + \bar{V}_{n+1}(x_{n+1},\tau_n,\mu_n)\},$$

in which the supremum is used because there may exist multiple inputs that maximize the value function.

Theorem 1 (Structure of the Optimal Policy) *Let* $D_n(\mu_n, \tau_n) \triangleq x^\star_{n+1}(\mu_n, \tau_n) - \mu_n/(2\alpha)$ *be the deviation of the optimal input,* $x^\star_{n+1}(\mu_n, \tau_n)$*, from the myopic input,* $\mu_n/(2\alpha)$*, we have*

1. *If* $\mu_n \geqslant 0$*, then* $D_n(\mu_n, \tau_n) \geqslant 0$.
2. *If* $\mu_n < 0$*, then* $D_n(\mu_n, \tau_n) < 0$.

Proof: We first consider the case of $\mu_n \geqslant 0$ and show that it is never optimal to choose $x_{n+1} < \mu_n/(2\alpha)$. More specifically, we show that, for any $x_{n+1} < \mu_n/(2\alpha)$, there exists a corresponding input $x'_{n+1} > \mu_n/(2\alpha)$ that yields a higher value. We analyze the cases of $x_{n+1} \geqslant 0$ and $x_{n+1} < 0$ separately below.

1. **Case I:** $x_{n+1} \geqslant 0$. To begin with, consider any $0 \leqslant x_{n+1} < \mu_n/(2\alpha)$, we can always find another input, $x'_{n+1} = \mu_n/(2\alpha) + [\mu_n/(2\alpha) - x_{n+1}] > \mu_n/(2\alpha) > x_{n+1}$, such that $-\alpha(x'_{n+1})^2 + \mu_n x'_{n+1} + \bar{V}_{n+1}(x'_{n+1}, \tau_n, \mu_n) = -\alpha x^2_{n+1} + \mu_n x_{n+1} + \bar{V}_{n+1}(x'_{n+1}, \tau_n, \mu_n) \geqslant -\alpha x^2_{n+1} + \mu_n x_{n+1} + \bar{V}_{n+1}(x_{n+1}, \tau_n, \mu_n)$, where the inequality follows from Proposition 4, namely, $\bar{V}_{n+1}(x_{n+1}, \tau_n, \mu_n)$ is increasing in x_{n+1} when $x_{n+1} > 0$. That is, x'_{n+1} yields a higher value than x_{n+1}, which hence cannot be optimal, see Figure 1.
2. **Case II:** $x_{n+1} < 0$. Now consider any $x_{n+1} < 0$, we can always find another input, $x'_{n+1} = \mu_n/(2\alpha) + [\mu_n/(2\alpha) - x_{n+1}] > \mu_n/(2\alpha) \geqslant 0 > x_{n+1}$, such that $-\alpha(x'_{n+1})^2 + \mu_n x'_{n+1} + \bar{V}_{n+1}(x'_{n+1}, \tau_n, \mu_n) = -\alpha x^2_{n+1} + \mu_n x_{n+1} + \bar{V}_{n+1}(x'_{n+1}, \tau_n, \mu_n) \geqslant -\alpha x^2_{n+1} + \mu_n x_{n+1} + \bar{V}_{n+1}(-x_{n+1}, \tau_n, \mu_n) = -\alpha x^2_{n+1} + \mu_n x_{n+1} + \bar{V}_{n+1}(x_{n+1}, \tau_n, \mu_n)$, where the inequality follows from Proposition 4, namely, $\bar{V}_{n+1}(x_{n+1}, \tau_n, \mu_n)$ is increasing in x_{n+1} when $x_{n+1} > 0$. This is because, $x'_{n+1} > -x_{n+1} \geqslant 0$ (also see Figure 1). The last equality follows from $\bar{V}_{n+1}(x_{n+1}, \tau_n, \mu_n) = \bar{V}_{n+1}(-x_{n+1}, \tau_n, \mu_n)$ in Proposition 4.

Therefore, the optimal input $x^\star_{n+1}(\mu_n, \tau_n)$ must be greater than or equal to $\mu_n/(2\alpha)$ when $\mu_n \geqslant 0$. This suggests that $D_n(\mu_n, \tau_n) \geqslant 0$ when $\mu_n \geqslant 0$ The case of $\mu_n < 0$ can be proved using the same argument, thus it is omitted here. *QED.*

References

[1] Bertsekas, D. (1976) *Dynamic programming and stochastic control.* Academic Press.

[2] Gittins, J. C. (1979) Bandit processes and dynamic allocation indices, *J. R. Stat. Soc. Ser. B*, 14: 148-177.

E-mail address: jw171@queensu.ca