Active monotonic regression

Abstract

"Forgetting functions" track the amount of information that a person remembers about an experience or event as it falls over time. The shape of the forgetting function provides a signature of the underlying memory systems responsible for the information's retention and loss. However, these functions are hard to measure accurately, typically requiring many thousands of trials. This precludes detailed study of the factors affecting the timecourse of forgetting and how it varies across individuals. Here, we present an adaptive procedure for measuring forgetting functions using active monotonic regression. By exploiting the structure inherent to the process of forgetting, we adaptively select the stimulus that is maximally informative about the forgetting function's shape. Using the adaptive procedure, we acheive a x-fold speedup in a standard working memory task and run the first large-scale experiment comparing individual differences in the rate of forgetting.

1 Introduction

There is a long tradition in psychology, one that goes back to Ebbinghaus's intial foray into the scientific study of memory, of describing the timecourse of remembering and forgetting using *forgetting functions*, which track the amount of information that a person remembers about an experience or event as it falls over time [3, 8]. Initially, much of what is seen or heard is remembered and accessible. Quickly at first, and then later more slowly, memory for the experience becomes degraded or inaccessible.

The form of the forgetting function provides a signature of the underlying memory systems responsible for the information's retention and the processes that lead to its degradation and loss over time [7]. For example, examining the form of the forgetting function for visual memory provided some of the primary evidence for the existence of iconic memory, a high-capacity sensory buffer [5]. The form of the forgetting function has also been instrumental to rational theories of adaptive forgetting [1, 2].

However, forgetting functions are hard to measure, typically requiring many thousands of trials for accurate measurement. The traditional method for measuring them selects a fixed set of evenly-spaced durations that span the relevant timescale, and then places an equal number of trials at each of these timepoints. When many forgetting functions are measured at once (as is the case, for example, when one wishes to compare the rate of forgetting across multiple experimental conditions), the functions are considered independently and an equal number of trials are allotted to each condition. With k conditions, m timepoints, and n trials per timepoint, kmn trials are needed to measure a single person's forgetting functions. This precludes detailed study of the factors affecting the timecourse of forgetting and how it varies across individuals.

The key to the ineffciency of the traditional approach is that it ignores the structure inherent to the form of the forgetting function, treating each data point as though it were equally valuable for determining the function's form. One example of this structure is that, due to the nature of forgetting, forgetting functions are inevitably monotonically decreasing — i.e., once a memory has been formed, and the original experience is no longer accessible, information about the experience is lost, not gained.

Here, we present an adaptive procedure for measuring forgetting functions using active monotonic regression. The procedure adaptively selects the stimulus that is expected to be maximally informative about the forgetting function's form. Using the adaptive procedure, we acheive a x-fold speedup in a standard working memory task and run the first large-scale experiment comparing individual differences in the rate of forgetting.

2 Adaptive procedures and active sampling

An active learning procedure is one in which the learner decides which data to access and observe [4]. One flavor of active learning is found in the sequential estimation procedures developed by psychophysicists to measure *threshold* — the stimulus intensity that separates the seen from the unseen and the heard from the unheard. The method of constant stimuli, discussed in some of the earliest work on perception, places an equal number of trials at each of a fixed set of stimulus intensities, and then uses the shape of the resulting psychometric function to estimate threshold. Methods such as QUEST and FAST improve on this technique by placing each trial at a stimulus intensity that is maximally informative for estimating threshold [6, **?**].

[Some background on active learning procedure relevant to our particular approach.]

3 Active monotonic regression

Here, we develop an active learning procedure for estimating forgetting functions under the assumption of a high-threshold model, where an experience is either entirely remembered of forgotten. [Kevin-note: be sure to reference Carpentier, Lazaric, et al. UCB algorithm for the variance balance problem.]

3.1 Problem description

Let $f : [0, 1] \to \mathbf{R}$ be an unknown function with the following known properties:

1. f **is bounded** : for any $0 \leq t \leq 1$ we have $0 \leq f(t) \leq 1$,
2. f **is non-increasing** : For any $0 \leq t \leq t' \leq 1$ we have that $f(t) \geq f(t')$,
3. f' **exists and is non-decreasing** : For any $0 \leq t \leq t' \leq 1$ we have that $f'(t) \leq f'(t') \leq 0$.

We say that $f \in \mathcal{F}$ if f obeys the above properties, and we are searching for a particular $f^* \in \mathcal{F}$. For any $t \in [0, 1]$ we may observe an independent Bernoulli random variable $Y_t \in \{0, 1\}$ where

$\mathbb{E}[Y_t] = f^*(t)$. Given a budget of m measurements, sequentially choose the location of these measurements in $[0,1]$ and use the observations to output a function $\widehat{f}_m(t)$ such that

$$||f^* - \widehat{f}_m||_2^2 = \int_0^1 |f(t) - \widehat{f}_m(t)|^2 dt \tag{1}$$

is small. With limited loss of generality, we study a discretized version of this problem where the coarseness of the discretization can be made arbitrarily fine. This discretization is performed only to have a computationally-efficient algorithm for finding $\widehat{f}_m$.

3.2 Reduction to finite dimensions

For some fixed n, define $x_i = \frac{i}{n}$ for $i = 0, 1, \ldots, n$ and $f^y(x) = \sum_{i=1}^n y_i \, \mathbf{1}\{x_{i-1} < x \leq x_i\}$ for some $y_i \in [0,1]$ for all $i = 1, \ldots, n$. If

$$S = \left\{ y \in \mathbb{R}^n : \; 0 \leq y_i \leq 1, \; y_j \leq y_i, \frac{y_j - y_i}{x_j - x_i} \leq \frac{y_k - y_j}{x_k - x_j} \quad \text{for all } 1 \leq i \leq j \leq k \leq n \right\}$$

where each set of inequalities is associated with the three defined conditions on our functions, then $\mathcal{F}_n := \{f^y : y \in S\} \subset \mathcal{F}$ and $\sup\limits_{f \in \mathcal{F}} \inf\limits_{f^y \in \mathcal{F}_n} = ||f - f^y||_2^2 \leq \dfrac{1}{n}$. Moreover, if $\mu := \arg\min\limits_{y \in S} ||f^* - f^y||_2$ then

$$\sup_{f^y \in \mathcal{F}_n} ||f^y - f^\mu||_2^2 = \int_0^1 |f^y(t) - f^\mu(t)|^2 dt = \sum_{i=1}^n |y_i - \mu_i|^2 \, |x_i - x_{i-1}| = \frac{1}{n} ||y - \mu||_2^2$$

and

$$\mathbb{E}||f^* - \widehat{f}_m||_2 \leq ||f^* - f^\mu|| + \mathbb{E}||f^\mu - \widehat{f}_m||_2 \leq \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n}} \mathbb{E}||\widehat{\mu}_m - \mu||_2 \tag{2}$$

where $\widehat{\mu}_m$ is our estimate of μ. Thus, the functions in $\mathcal{F}_n$ faithfully represent the functions of $\mathcal{F}$ arbitrarily well as $n \to \infty$.

Note that for a fixed n, if S has non-zero volume and $\mu \in S \setminus \partial S$ then there exists an m_0 such that $\mathbb{E}||\widehat{\mu}_m - \mu||_2 = \Theta(\sqrt{\frac{n^2}{m}})$ for all $m \geq m_0$. As a learning rate with respect to m, this parametric rate is only possible because we have essentially made the problem parametric by fixing n. Inspecting (2), one realizes that to achieve the expected nonparametric rates of convergence for $\widehat{f}_m$, we must have that n increases with m. A natural choice is to set $n = m^p$ where $p \in (0,1]$ is chosen to match the true learning rate[1]. The intuition is that for n growing this fast with m, any $\mu \in S$ is very close to a point in ∂S and the geometry of the tangent cone at this point is what controls the learning rate. However, practical constraints sometimes make an unbounded growing n impossible or undesirable. In what follows, we will consider a fixed n and then later describe how to grow n with m in order to achieve the desired learning rates.

[1] In practice, for arbitrary sets S this information will not be known. We overcome this difficulty by devising strategies that adaptively set n once some amount of data has been collected.

4 Estimation and measurement allocations

Let ν be a single parameter exponential family parameterized by its mean, e.g. $\mathbb{E}_{\nu(\theta)}[X] = \theta$. Assume each observation from the ith arm is drawn i.i.d. from the distribution $\nu(\mu_i)$. Suppose we were given τ_i observations from each arm i and we wish to provide an estimate $\widehat{\mu}$ of μ in terms of some given norm $||\cdot||$. A natural estimate would be to choose the $\widehat{\mu} \in K$ that maximizes the likelihood of observing the data:

$$\max_{\theta \in K} \prod_{i=1}^{n} \prod_{j=1}^{\tau_i} d\nu(X_{i,j}; \theta_i) \tag{3}$$

For example, if $\nu(\theta) = \mathcal{N}(\theta, 1)$ then the maximum likelihood estimator (MLE) of (3) is equivalent to

$$\begin{aligned} \arg\max_{\theta \in K} \prod_{i=1}^{n} \prod_{j=1}^{\tau_i} \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}(X_{i,j} - \theta_i)^2\right) &= \arg\min_{\theta \in K} \sum_{i=1}^{n} \sum_{j=1}^{\tau_i} (X_{i,j} - \theta_i)^2 \\ &= \arg\min_{\theta \in K} \sum_{i=1}^{n} \tau_i (\widetilde{\mu}_i - \theta_i)^2 \\ &= \arg\min_{\theta \in K} \sum_{i=1}^{n} \tau_i KL(\widetilde{\mu}_i, \theta_i) \end{aligned}$$

and if $\nu(\theta) = \text{Bernoulli}(\theta)$ then

$$\begin{aligned} \arg\max_{\theta \in K} \prod_{i=1}^{n} \prod_{j=1}^{\tau_i} \theta_i^{X_{i,j}} (1-\theta_i)^{1-X_{i,j}} &= \arg\max_{\theta \in K} \sum_{i=1}^{n} \sum_{j=1}^{\tau_i} X_{i,j} \log(\theta_i) + (1 - X_{i,j}) \log(1-\theta_i) \\ &= \arg\max_{\theta \in K} \sum_{i=1}^{n} \tau_i [\widetilde{\mu}_i \log(\theta_i) + (1-\widetilde{\mu}_i) \log(1-\theta_i)] \\ &= \arg\min_{\theta \in K} \sum_{i=1}^{n} \tau_i \left[\widetilde{\mu}_i \log(\frac{\widetilde{\mu}_i}{\theta_i}) + (1-\widetilde{\mu}_i) \log(\frac{1-\widetilde{\mu}_i}{1-\theta_i}) \right] \\ &= \arg\min_{\theta \in K} \sum_{i=1}^{n} \tau_i KL(\widetilde{\mu}_i, \theta_i). \end{aligned}$$

In what follows, for whatever ν is under consideration, we will use the estimator

$$\widehat{\mu} = \arg\min_{\theta \in K} \sum_{i=1}^{n} \tau_i KL(\widetilde{\mu}_i, \theta_i). \tag{4}$$

4.1 PAC measurement allocation

If we are given μ (or equivalently $\nu(\mu_i)$) for each i then for any $\tau \in \mathbb{N}^n$ we can simulate τ_i observations from $\nu(\mu_i)$ for each i, construct $\widehat{\mu}_\tau$, and compute $||\widehat{\mu}_\tau - \mu||$. We can also repeat this process m times (with the same fixed τ) to obtain estimates $\widehat{\mu}_\tau^{(1)}, \dots, \widehat{\mu}_\tau^{(m)}$ and start to estimate the quantity $\mathbb{P}\left(||\widehat{\mu}_\tau - \mu|| \geq \epsilon\right)$ for some fixed ϵ. Note that this probability is a deterministic function of τ and for a given fixed ϵ, δ we can define the optimization program

$$\begin{aligned} \underset{\tau \in \mathbb{R}_+^n}{\text{minimize}} \quad & \sum_{i=1}^n \tau_i \\ \text{subject to} \quad & \mathbb{P}\left(||\widehat{\mu}_\tau - \mu|| \geq \epsilon\right) \leq \delta. \end{aligned} \tag{5}$$

While estimating $\mathbb{P}\left(||\widehat{\mu}_\tau - \mu|| \geq \epsilon\right)$ is theoretically possible, in practice this can be quite time-consuming as the computation is at least $\Omega(\delta^{-1})$. In this work we focus on the ℓ_∞ and ℓ_2 norms and for these cases we propose analyses and algorithms that are more efficient than the most general case.

4.2 Max-width minimization for ℓ_∞ objective

4.2.1 Oracle allocation

In practice we can use a standard Chernoff bound to provide subsets such that for any fixed τ, δ we have $\mathbb{P}\left(\widehat{\mu}_\tau - \mu \in V(\tau)\right) \geq 1 - \delta$ where $V(\tau) \subset \mathbb{R}^n$. This allows us to restate (5) as

$$\begin{aligned} \underset{\tau \in \mathbb{R}_+^n}{\text{minimize}} \quad & \sum_{i=1}^n \tau_i \\ \text{subject to} \quad & ||y - \mu|| \leq \epsilon, \; \forall y \in K \cap (\mu + V(\tau)). \end{aligned} \tag{6}$$

that is sufficient to guarantee $\mathbb{P}\left(||\widehat{\mu}_\tau - \mu|| \geq \epsilon\right) \geq 1 - \delta$.

Recall that we wish to identify this somehow optimal τ so that if we used this allocation to take measurements and to build an estimate for $\widehat{\mu}_\tau$, we have a guarantee that this number of measurements is minimal with respect to all observation vectors τ that guarantee $\mathbb{P}\left(||\widehat{\mu}_\tau - \mu|| \geq \epsilon\right) \geq 1-\delta$. But, of course, identifying this optimal τ requires knowledge of μ, which is what we are trying to estimate in the first place! This paper explores methods of adapting the sampling distribution over time as data is observed so that as time went on the empirical sampling distribution would approach this optimal τ.

Consider a rewritten version of the problem of (6):

$$\begin{aligned} \underset{\tau \in \mathbb{R}_+^n}{\text{minimize}} \quad & \sum_{i=1}^n \tau_i \\ \text{subject to} \quad & ||x||_\infty \leq \epsilon, \;\; \forall x \in (K - \mu) \cap V(\tau). \end{aligned} \tag{7}$$

Lemma 1 *Let* (τ, y) *and* (τ, x) *be the solutions to the optimization problems* (6) *and* (7)*, respectively. Then* $||y - \mu|| = ||x|| = \inf_{u:||u||=1} \sup_{t \in V_i} \langle u, t \rangle = \epsilon$*. If* $\tau \mapsto \alpha^2 \tau$ *then* $\rho \mapsto \rho/\alpha$*. Also,* $\alpha V(\alpha^2 \tau) \subseteq V(\tau)$*.*

4.2.2 Adaptive allocation

Consider the following optimization programs that do not require μ

$$\begin{aligned} \underset{\tau \in \mathbb{R}^n_+}{\text{minimize}} \quad & \sum_{i=1}^{n} \tau_i \\ \text{subject to} \quad & ||y - \nu||_\infty \leq \alpha\rho, \quad \forall y \in K \cap (\nu + \alpha V(\tau)), \\ & \qquad\qquad\qquad\quad \forall \nu \in K \cap (\widehat{\mu} + V(\tau^-)). \end{aligned} \tag{8}$$

$$\begin{aligned} \underset{\tau \in \mathbb{R}^n_+}{\text{minimize}} \quad & \sum_{i=1}^{n} \tau_i \\ \text{subject to} \quad & ||x||_\infty \leq \alpha\rho - \kappa, \ \forall x \in (K - \widehat{\mu}) \cap \alpha V(\tau). \end{aligned} \tag{9}$$

$$\begin{aligned} \underset{\tau \in \mathbb{R}^n_+}{\text{minimize}} \quad & \sum_{i=1}^{n} \tau_i \\ \text{subject to} \quad & ||y - \widehat{\mu}||_\infty \leq \alpha\rho - \kappa, \ \forall x \in K \cap \widehat{\mu} + \alpha V(\tau). \end{aligned} \tag{10}$$

$$\begin{aligned} \underset{\tau \in \mathbb{R}^n_+}{\text{minimize}} \quad & \sum_{i=1}^{n} \tau_i \\ \text{subject to} \quad & ||x + \mu - \widehat{\mu}||_\infty \leq \alpha\rho - \kappa, \ \forall x \in (K - \mu) \cap \widehat{\mu} - \mu + \alpha V(\tau). \end{aligned} \tag{11}$$

Lemma 2 *Fix α, ρ and assume that $\sup_{y \in K \cap (\widehat{\mu} + V(\tau^-))} ||y - \mu|| \leq \kappa < \alpha\rho$. Let $K^o_\mu = \{x \in \mathbb{R}^n : \exists \eta > 0 : \eta(x - \mu) \in K - \mu\}$ so that $K^o_\mu - \mu$ is a cone. Suppose $(K^o_\mu - \mu) \cap (\alpha\rho + \kappa)[-1,1]^n = (K - \mu) \cap (\alpha\rho + \kappa)[-1,1]^n$. Let (τ, y) and (τ^*, y^*) be (possible non-unique) solutions to the optimization problems of* (8) *and* (6)*, respectively. Then $\sup_{y \in K \cap (\mu + V(\tau))} ||y - \mu|| \leq \rho$ and $\sum_{i=1}^n \tau_i \leq (1 - \frac{\kappa}{\alpha\rho})^{-2} \sum_{i=1}^n \tau_i^*$.*

Proof Because the optimization of $y - \nu$ will always be contained within $(\alpha\rho + \kappa)[-1,1]^n$ we may assume that $K - \mu$ is a cone since in this region $K - \mu = K^o_\mu - \mu$ so that $\alpha(K - \mu) = (K - \mu)$ for any $\alpha > 0$. We begin by rewriting the optimization problem as follows:

$$\begin{aligned} \underset{\tau \in \mathbb{R}^n_+}{\text{minimize}} \quad & \sum_{i=1}^{n} \tau_i \\ \text{subject to} \quad & ||x - \epsilon|| \leq \alpha\rho, \quad \forall x \in \alpha(K - \mu) \cap (\epsilon + \alpha V(\tau)), \\ & \qquad\qquad\qquad\ \forall \epsilon \in (K - \mu) \cap (\widehat{\mu} - \mu + V(\tau^-)). \end{aligned} \tag{12}$$

Thus, for any α, τ, ϵ

$$\begin{aligned} \{x : ||x - \epsilon|| \leq & \alpha\rho, \ x \in \alpha(K - \mu) \cap (\epsilon + \alpha V(\tau))\} \\ & = \{x : ||x - \epsilon/\alpha|| \leq \rho, \ x \in (K - \mu) \cap (\epsilon/\alpha + V(\tau))\}. \end{aligned}$$

The τ solution of (8) is minimal, so it suffices to show any allocation of τ that satisfies the constraints that is not much more than the optimal allocation τ^* of (6) if μ was known. In particular, we can start with τ^* and construct a new τ. From (1), if we begin with a particular τ^* that achieves a particular ρ, then if we use $\gamma^{-2}\tau^*$ we achieve $\rho\gamma$. Since $||\epsilon|| \leq \kappa$, it suffices to bring ρ down to $\rho - \kappa/\alpha$, resulting in a $\gamma = (1 - \frac{\kappa}{\alpha\rho})$. We conclude that if τ is the solution of (8) then $\sum_{i=1}^n \tau_i \leq (1 - \frac{\kappa}{\alpha\rho})^{-2} \sum_{i=1}^n \tau_i^*$ ■

Until S starts to look like a cone (as has been assumed about K), the procedure is oblivious to the structure of S in the worst case, requiring $O(n\rho^{-2})$ measurements per round in the worst case. S starts "looking" like a cone as soon as $\{x : (S - \mu) \cap [-\alpha\rho, \alpha\rho]^n\}$ starts looking like truncated cone. This explains why we cannot simply take α to be arbitrarily large.

The algorithm is quite simple: On the kth round, define $\rho = 2^{-k}$ so on each round the sub-optimality constant is equal to $(1 - 2/\alpha)^{-2}$ for any $\alpha > 2$.

4.2.3 Solving the optimization problem of (8)

The optimization problem of (8) is non-convex in general which means that obtaining a global optimal may not be possible. However, local search methods are possible.

For general norms, computing $\sup\limits_{x \in (K-\mu)\cap V(\tau)} ||x||$ is computationally intractable. But at least for $||\cdot||_\infty$ computing this so-called max-width is computable using a number of linear programs. This motivates the following algorithm:

$$\mathcal{L}(\tau, \lambda) = \sum_{i=1}^n \tau_i + \lambda \left(\sup_{x \in (K-\mu)\cap V(\tau)} ||x|| - \rho \right)$$

Define $f(\tau) = \sup\limits_{x \in (K-\mu)\cap V(\tau)} ||x||$ and the approximate derivative:

$$g_i^k = 1 + \lambda^k \frac{f(\tau^k + \gamma \mathbf{e}_i) - f(\tau^k)}{\gamma}$$

which allows us to define the following algorithm:

$$\begin{aligned} \tau^{k+1} &= \left[\tau^k - \eta\, g^k\right]_+ \\ \lambda^{k+1} &= \left[\lambda^k + \eta\, \left(f(\tau^{k+1}) - \rho\right)\right]_+ \end{aligned}$$

4.2.4 Constructing confidence V subsets

Consider when random variables are Bernoulli versus Gaussian, how do things change?

4.3 Mean-width minimization for ℓ_2 objective

4.3.1 Projections onto Convex Sets

Let $I_1, \ldots, I_m$ index m observations taking values in $\{1, \ldots, n\}$ and suppose we observe the pairs $\{(x_{I_i}, Y_{I_i})\}_{i=1}^m$ where $Y_{I_i} \sim \nu(\mu_{I_i})$ as before for all $i = 1, \ldots, m$. For convenience let $A = (\mathbf{e}_{I_1}^T, \ldots, \mathbf{e}_{I_m}^T)^T$ and $b = (Y_{I_1}, \ldots, Y_{I_m})^T$. The next lemma contains a number of simple but useful identities used throughout the analysis.

Lemma 3 *Let K be an arbitrary convex set in $\mathbb{R}^n$. If $\Pi_K(x) = \arg\min_{\nu \in K} ||\nu - x||_2^2$, $\Pi_{K,H}(x) = H^{-1/2}\Pi_{H^{1/2}K}(H^{1/2}x)$ and*

$$\widetilde{\mu} = \arg\min_{\nu \in \mathbb{R}^n} \sum_{i=1}^m (\nu_{I_i} - Y_{I_i})^2 = \arg\min_{\nu \in \mathbb{R}^n} ||A\nu - b||_2^2$$

then the following are all equal

1. $\arg\min_{\nu \in K} \sum_{i=1}^m (\nu_{I_i} - Y_{I_i})^2$
2. $\arg\min_{\nu \in K} ||A\nu - b||_2^2$
3. $\arg\min_{\nu \in K} ||\nu - \widetilde{\mu}||_{A^TA}^2$
4. $\Pi_{K,A^TA}(\widetilde{\mu})$.

Proof By calculus, $\widetilde{\mu} = (A^TA)^{-1}A^Tb$. We will show (4) = (3) = (2).

$$\begin{aligned}
\Pi_{S,A^TA}(\widetilde{\mu}) &= (A^TA)^{-1/2}\Pi_{(A^TA)^{1/2}S}((A^TA)^{1/2}\widetilde{\mu}) \\
&= (A^TA)^{-1/2} \arg\inf_{\nu \in (A^TA)^{1/2}S} ||\nu - (A^TA)^{1/2}\widetilde{\mu}||_2^2 \\
&= \arg\inf_{\nu \in K} ||(A^TA)^{1/2}\nu - (A^TA)^{1/2}\widetilde{\mu}||_2^2 \\
&= \arg\inf_{\nu \in K} ||\nu - \widetilde{\mu}||_{A^TA}^2 \qquad (4) = (3) \\
&= \arg\inf_{\nu \in K} \nu^TA^TA\nu - 2\nu^TA^TA\widetilde{\mu} + \widetilde{\mu}^T\widetilde{\mu} \\
&= \arg\inf_{\nu \in K} \nu^TA^TA\nu - 2\nu^TA^Tb + b^Tb \\
&= \arg\min_{\nu \in K} ||A\nu - b||_2^2 \qquad (3) = (2) \\
&= \arg\min_{\nu \in K} \sum_{i=1}^m (\mathbf{e}_{I_i}^T\nu - Y_{I_i})^2 \\
&= \arg\min_{\nu \in K} \sum_{i=1}^m (\nu_{I_i} - Y_{I_i})^2 \qquad (2) = (1)
\end{aligned}$$

which completes the proof. ■

The above lemma suggests the following procedure after observing $\{(x_{I_i}, Y_{I_i})\}_{i=1}^m$: compute $\widetilde{\mu} \in \mathbb{R}^n$ and use (3) to solve for $\widehat{\mu} \in K$ where in the sequel, K will be a subset of S. The next lemma will give us incite into the quality of this estimate.

Lemma 4 (Plan, Vershynin, Yudovina 2014) *Let A be defined as above, let K be a star-shaped*[2] *set and let $z \in K$, $w \in \mathbb{R}^n$. Then for every $t > 0$ we have*

$$\begin{aligned} ||\Pi_{K,A^TA}(w) - z||_{A^TA} &\leq \max\left\{t, \frac{2}{t} \sup_{x \in K: ||x-z||_{A^TA} \leq t} \langle x - z, w - z\rangle_{A^TA}\right\} \\ &= \max\left\{t, \frac{2}{t} \sup_{u \in (A^TA)^{1/2}(K-z): ||u||_2 \leq t} \langle u, (A^TA)^{1/2}(w - z)\rangle\right\} \\ &\leq \sqrt{2 \sup_{u \in (A^TA)^{1/2}(K-z)} \langle u, (A^TA)^{1/2}(w-z)\rangle}. \end{aligned}$$

Moreover, $||\Pi_{K,A^TA}(w) - z||_2 \leq ||\Pi_{K,A^TA}(w) - z||_{A^TA} / \sqrt{\min_i \mathbf{e}_i^T A^T A \mathbf{e}_i}$.

Proof Let $d = ||\Pi_{K,A^TA}(w) - z||_{A^TA}$. By Lemma 3 $\Pi_{K,A^TA}(w)$ is the closest vector in K under the A^TA metric so

$$||\Pi_{K,A^TA}(w) - w||_{A^TA} \leq ||z - w||_{A^TA}.$$

By squaring both sides, subtracting and adding z within the first term, expanding, and simplifying, gives

$$||\Pi_{K,A^TA}(w) - z||_{A^TA}^2 \leq 2\langle \Pi_{K,A^TA}(w) - z, w - z\rangle_{A^TA}.$$

The left hand side is equal to d^2. By definition, both $\Pi_{K,A^TA}(w)$ and z live in K and are separated by a distance d under the A^TA metric, so

$$d^2 \leq 2 \sup_{x \in K: ||x-z||_{A^TA} \leq d} \langle x - z, w - z\rangle_{A^TA}.$$

To prove the lemma, if $d \leq t$, the lemma holds. If $d > t$ then by the above display

$$d \leq \frac{2}{d} \sup_{x \in K: ||x-z||_{A^TA} \leq d} \langle x - z, w - z\rangle_{A^TA} \leq \frac{2}{t} \sup_{x \in K: ||x-z||_{A^TA} \leq t} \langle x - z, w - z\rangle_{A^TA}.$$

■

An important special case of the lemma is when $\nu = \mathcal{N}(0,1)$ because it transforms the stochastic estimation problem into a purely geometric one. It is useful to introduce the notion of the Gaussian width of a set K which describes the "complexity" of the set in some sense.

[2] A set K is star-shaped if there exists a point $x_0 \in K$ such that for all $x \in K$, the line connecting x and x_0 is contained within K. If K is convex then K is star-shaped.

Definition 1 (Gaussian Width) *Fix some set $K \subseteq \mathbb{R}^n$. The* global Gaussian width *of a subset $K \subset \mathbb{R}^n$ is defined as*

$$W(K) = \mathbb{E}\left[\sup_{x \in K} \langle x, Z \rangle\right]$$

where $Z \sim \mathcal{N}(0, I)$. The local Gaussian width *of a subset $K \subset \mathbb{R}^n$ is a function of scale $t \geq 0$, and it is defined as*

$$W_t(K) = \mathbb{E}\left[\sup_{x \in K : ||x||_2 \leq t} \langle x, Z \rangle\right].$$

Lemma 5 *Let A and b be defined as before. If $Y_{I_i} = \mu_{I_i} + \xi_i$ where $\xi_i \sim \mathcal{N}(0, 1)$ then for any $t \geq 0$ we have with probability at least $1 - \delta$*

$$\begin{aligned} ||\Pi_{K,A^TA}(\widetilde{\mu}) - \mu||_{A^TA} &\leq \max\left\{t, \frac{2W_t((A^TA)^{1/2}(K-\mu))}{t}\right\} + \sqrt{8\log(1/\delta)} \\ &\leq \sqrt{2W((A^TA)^{1/2}(K-\mu))} + \sqrt{8\log(1/\delta)}. \end{aligned}$$

Proof Recall that $\widetilde{\mu} = \arg\min_{\nu \in \mathbb{R}^n} ||A\nu - b||_2^2 = (A^TA)^{-1}A^Tb$ so that

$$\begin{aligned} \widetilde{\mu} - \mu &= (A^TA)^{-1}A^Tb - \mu \\ &= (A^TA)^{-1}A^TA\mu + (A^TA)^{-1}A^T\xi - \mu \\ &= (A^TA)^{-1}A^T\xi. \end{aligned}$$

We note that

$$\mathbb{E}[(A^TA)^{-1}A^T\xi\xi^TA(A^TA)^{-1}] = (A^TA)^{-1}$$

which means that $(A^TA)^{-1}A^T\xi = (A^TA)^{-1/2}Z$ in distribution where $Z \sim \mathcal{N}(0, I)$. If we set $w = \widetilde{\mu}$ and $z = \mu$ in the above lemma, we observe that

$$\begin{aligned} &\mathbb{E}\left[\sup_{u \in (A^TA)^{1/2}(K-\mu) : ||u||_2 \leq t} \langle u, (A^TA)^{1/2}(\widetilde{\mu} - \mu)\rangle\right] \\ &\qquad = \mathbb{E}\left[\sup_{u \in (A^TA)^{1/2}(K-\mu) : ||u||_2 \leq t} \langle u, Z \rangle\right] = W_t((A^TA)^{1/2}(K-\mu)) \end{aligned}$$

which explains the first term. The second term describes the deviation from the mean which we explore next. Note that for any Z, $\mathbb{E}[\langle u, Z\rangle] = 0$, $\mathbb{E}[\langle u, Z\rangle^2] = ||u||^2 \leq t^2$ for all $u \in (A^TA)^{1/2}(K-\mu) : ||u||_2 \leq t$. Thus, by applying a standard sub-Gaussian tail bound to the suprema we obtain the claimed result. ■

Note that for any $K \subset \mathbb{R}^n$ and any $t > 0$, $W_t(K)$ can be computed to arbitrary accuracy through stochastic simulation. Moreover, using the fact that $\frac{W_t(K)}{t}$ is monotonically decreasing in t, we realize that $\max\{t, \frac{2W_t(K)}{t}\}$ is pseudo-convex and is as simple as a one-dimensional line-search, e.g. Golden-section search.

4.3.2 Oracle Allocation Strategy

Now recall that we are searching over allocations such that we observe the ith component τ_i times. This translates into the following optimization program given the above preliminaries:

$$\begin{aligned} \underset{\tau \in \mathbb{R}^n_+}{\text{minimize}} \quad & \sum_{i=1}^n \tau_i \\ \text{subject to} \quad & \frac{\inf_t \max\{t, \frac{2}{t} W_t\big(\text{diag}(\sqrt{\tau})(K-\mu)\big)\} + \sqrt{8\log(1/\delta)}}{\sqrt{\min_i \tau_i}} \leq \epsilon \end{aligned} \tag{13}$$

In this display we immediately gain insight into the fundamental hardness of the problem. As claimed above, if $\mu \in \text{interior}(K)$ and $\min_i \tau_i$ is large enough, $\text{diag}(\sqrt{\tau})(K-\mu)$ will start appearing like $\mathbb{R}^n$ and the best allocation for τ is uniform. In this regime we have that $\inf_t \max\{t, \frac{2}{t} W_t\big(\text{diag}(\sqrt{\tau})(K-\mu)\big)\} \approx \mathbb{E}||Z||_2 \approx \sqrt{n}$ and the asymptotic rate for ϵ acts like $\sqrt{\frac{n^2}{m}} + \sqrt{\frac{n\log(1/\delta)}{m}}$ which is exactly the rate of estimation when $K = \mathbb{R}^n$. Thus, learning rates as a function of m are only interesting for $\epsilon \gg 0$ or $\mu \in \partial K$.

4.3.3 Adaptive Allocation Strategy

$$\begin{aligned} \underset{\tau \in \mathbb{R}^n_+}{\text{minimize}} \quad & \sum_{i=1}^n \tau_i \\ \text{subject to} \quad & \frac{\inf_t \max\{t, \frac{2\alpha}{t} W_t\big(\text{diag}(\sqrt{\tau\alpha^{-2}})(K-\widehat{\mu})\big)\} + \sqrt{8\log(1/\delta)}}{\sqrt{\min_i \tau_i}} \leq \epsilon - \kappa \end{aligned} \tag{14}$$

4.3.4 Brief aside: The Chatterjee method

Define

$$f_\mu(t) = \mathbb{E}\left[\sup_{\nu \in K : ||\nu-\mu||_2 \leq t} \langle Z, \nu - \mu \rangle\right] - \frac{t^2}{2}.$$

It turns out that if $\widetilde{\mu} = \mu + Z$ and $\widehat{\mu} = \arg\min_{\nu \in K} ||\widetilde{\mu} - \nu||_2$ then $||\mu - \widehat{\mu}||_2$ concentrates around $t_\mu = \arg\max_t f_\mu(t)$. Now consider some other μ':

$$\begin{aligned}
f_{\mu'}(t) &= \mathbb{E}\left[\sup_{\nu \in K : ||\nu - \mu'||_2 \le t} \langle Z, \nu - \mu' \rangle\right] - \frac{t^2}{2} \\
&= \mathbb{E}\left[\sup_{\nu \in K : ||\nu - \mu + \mu - \mu'||_2 \le t} \langle Z, \nu - \mu + \mu - \mu' \rangle\right] - \frac{t^2}{2} \\
&\le \mathbb{E}\left[\sup_{\nu \in K : ||\nu - \mu||_2 \le t + ||\mu - \mu'||_2} \langle Z, \nu - \mu \rangle\right] - \frac{t^2}{2} \\
&\le \mathbb{E}\left[\sup_{\nu \in K : ||\nu - \mu||_2 \le t + ||\mu - \mu'||_2} \langle Z, \nu - \mu \rangle\right] - \frac{(t+\Delta)^2}{2} + \frac{(t+\Delta)^2 - t^2}{2} \\
&= f_\mu(t+\Delta) + \frac{\Delta(2t+\Delta)}{2} \\
&\le f_\mu(t_\mu) - \frac{(t+\Delta-t_\mu)^2}{2} + \frac{\Delta(2t+\Delta)}{2} \\
&= f_\mu(t_\mu) - \frac{(t-t_\mu)^2}{2} + \Delta t_\mu .
\end{aligned}$$

It is likely the same tricks of above can be applied here to show that we can design A optimally by inflating the confidences and thereby shrinking Δ here.

5 Measuring individual difference in forgetting

In this section, we use active monotonic regression to reduce the time needed to measure a forgetting function to the point where it can be run in a largescale online experiment, allowing us to measure individual differences in the shape of the forgetting function.

6 Discussion

References

[1] J. R. Anderson. A rational analysis of human memory. In H. L. Roediger and F. I. M. Craik, editors, *Varieties of memory and consciousness: Essays in honour of Endel Tulving*, pages 195–210. 1989.

[2] John R Anderson and Lael J Schooler. Reflections of the environment in memory. *Psychological Science*, 2(6):396–408, 1991.

[3] Hermann Ebbinghaus. *Memory: A contribution to experimental psychology*. Teachers College, Columbia University, 1913.

[4] Todd M Gureckis and Douglas B Markant. Self-directed learning a cognitive and computational perspective. *Perspectives on Psychological Science*, 7(5):464–481, 2012.

[5] G Sperling. The information available in brief visual presentations. *Psychological Monographs*, 74, 1960.

[6] Andrew B Watson and Denis G Pelli. Quest: A bayesian adaptive psychometric method. *Perception & psychophysics*, 33(2):113–120, 1983.

[7] John T Wixted. Analyzing the empirical course of forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(5):927–935, 1990.

[8] John T Wixted and Ebbe B Ebbesen. On the form of forgetting. *Psychological Science*, 2(6):409–415, 1991.