# Contextual Learning with Online Convex Optimization: Theory and Application to Medical Decision-Making (Online Appendices)

Esmaeil Keyvanshokooh[1], Mohammad Zhalechian[2], Cong Shi[3], Mark P. Van Oyen[4], Pooyan Kazemian[5]
[1]Information and Operations Management, Mays Business School, Texas A&M University, College Station, TX 77845.
[2]Operations and Decision Technologies, Kelley School of Business, Indiana University, Bloomington, IN 47405.
[3]Management Science, Herbert Business School, University of Miami, Coral Gables, FL 33146.
[4]Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI 48105.
[5]Weatherhead School of Management, Case Western Reserve University, Cleveland, OH 44106.
keyvan@tamu.edu, mzhale@iu.edu, cxs2089@miami.edu, vanoyen@umich.edu, pxk409@case.edu.

## Appendix

### A. Technical Results on Contextual Learning Loss

PROPOSITION 4 **(Confidence Bound on Expected Reward under Stochastic Delays)**. *Let $D(t)$ be the stochastic delay time for observing the feedback of patient $t$, and assume that $\{D(t)\}_{t=1}^{T}$ are i.i.d. non-negative random variables with mean $\mu_D$ that satisfy the following regularity condition for any finite $m$:*

$$\mathbb{P}\big(D(t)-\mu_D \geq m\big) \leq \exp\Big(-\frac{m^{p+1}}{2\sigma_D^2}\Big),$$

*for some $p \geq 0$ and $\sigma_D > 0$, which characterize the tail of delay distribution.*

*Then, for any $t \geq t_0$ and $\delta > 0$ , where $t_0 = \max\{t_1, t_2\}$ in which $t_1$ and $t_2$ are defined in Lemma 2, the following estimation loss bound on the difference between the true and estimated expected rewards under each selected medication $k$ holds with probability at least $1-4\delta$:*

$$\Big|\mathbf{V}_t^{(k,y)}\big(\theta,\pi\big) - \mathbf{V}_t^{(k,y)}\big(\hat{\theta}(t),\hat{\pi}(t)\big)\Big| \leq \mathrm{Rad}_k(t),$$

*where $\hat{\theta}(t)$ and $\hat{\pi}(t)$ are respectively the maximum likelihood estimators of the unknown model parameters $\theta$ and $\pi$ at time period $t$ obtained under stochastic delays, and the confidence radius $\mathrm{Rad}_k(t)$ is defined as:*

$$\begin{aligned}\mathrm{Rad}_k(t) = \frac{1}{4}\left\|\phi_k(t)\right\|_{V_t^{-1}}\left(\frac{1}{c_\sigma}\sqrt{d_1\log\left(\frac{(t-N(t))\,c_\phi^2}{d_1}\right)+\log\left(\frac{1}{\delta^2}\right)}+c_\phi\sqrt{N(t)}\right)\\ +\frac{\nu}{2}\left\|\chi_k(t)\right\|_{U_t^{-1}}\left(\lambda\sqrt{(d_2+K)\log\left(\frac{(t-N(t))\,c_\chi^2}{d_2+K}\right)+\log\left(\frac{1}{\delta^2}\right)}+c_\chi\sqrt{N(t)}\right),\end{aligned}$$

*where $c_\sigma = \inf\big\{\nabla_{\theta_0}\sigma\big(\phi_k^\top\theta_0 - f_k(p_k;\pi_0)\big) : \|\theta_0-\theta\| \leq 1, \|\pi_0\| \leq 1, \|\phi_k\| \leq c_\phi, \|\chi_k\| \leq c_\chi\big\} > 0$ and also $\nu = \max\{\alpha_k,\beta_k\}/2$. Further, $V_t = \sum_{s=1}^{t-1}\phi_k(s)\,\phi_k(s)^\top$ and $U_t = \sum_{s=1}^{t-1}\chi_k(s)\,\chi_k(s)^\top$ are the design matrices corresponding to the first $t-1$ observed features. $N(t)$ is the random number of unrealized feedback outcomes at time period $t$, which is upper bounded by $2\mu_D + \tilde{\sigma}\sqrt{2\log\big(1/\delta\big)}+c$ with probability at least $1-\delta$, in which $c = 2\,\tilde{\sigma}^2\log\left(2\sigma_D^2+1\right)+1$ and $\tilde{\sigma} = \sigma_D\sqrt{p+2}$.*

*Proof of Proposition 4:* The objective is to establish a high probability confidence bound that contains the true expected reward using the feedback outcomes realized by the end of time period $t$.

To this aim, consider that medication $k$ is the one selected by the SGD-MAB algorithm for patient $t$. We first decompose $|\mathbf{V}_t^{(k,y)}(\theta,\pi) - \mathbf{V}_t^{(k,y)}(\hat{\theta}(t),\hat{\pi}(t))|$, which is the difference between the true and estimated expected rewards of patient $t$ under the selected medication $k$, by the triangle inequality as follows:

$$\begin{aligned}\left|\mathbf{V}_t^{(k,y)}\big(\theta,\pi\big) - \mathbf{V}_t^{(k,y)}\big(\hat{\theta}(t),\hat{\pi}(t)\big)\right| &= \left|\sigma\Big(\phi_k(t)^\top\theta - f_k\big(p_k(t);\pi\big)\Big) - \sigma\Big(\phi_k(t)^\top\hat{\theta}(t) - f_k\big(p_k(t);\hat{\pi}(t)\big)\Big)\right| \\ &\le \underbrace{\left|\sigma\Big(\phi_k(t)^\top\theta - f_k\big(p_k(t);\pi\big)\Big) - \sigma\Big(\phi_k(t)^\top\hat{\theta}(t) - f_k\big(p_k(t);\pi\big)\Big)\right|}_{\text{Term (I)}} \\ &\quad + \underbrace{\left|\sigma\Big(\phi_k(t)^\top\hat{\theta}(t) - f_k\big(p_k(t);\pi\big)\Big) - \sigma\Big(\phi_k(t)^\top\hat{\theta}(t) - f_k\big(p_k(t);\hat{\pi}(t)\big)\Big)\right|}_{\text{Term (II)}}, \end{aligned} \tag{9}$$

where $\hat{\theta}(t)$ and $\hat{\pi}(t)$ are the estimators of the unknown model parameters $\theta$ and $\pi$ at time period $t$, respectively.

**Term (I)**: To derive a high-probability bound for this term, we first decompose it as follows:

$$\begin{aligned}&\left|\sigma\Big(\phi_k(t)^\top\theta - f_k\big(p_k(t);\pi\big)\Big) - \sigma\Big(\phi_k(t)^\top\hat{\theta}(t) - f_k\big(p_k(t);\pi\big)\Big)\right| \\ &\le \frac{1}{4}\left|\phi_k(t)^\top\theta - \phi_k(t)^\top\hat{\theta}(t)\right| = \frac{1}{4}\left|\Big[V_t^{1/2}\big(\theta-\hat{\theta}(t)\big)\Big]^\top\Big[V_t^{-1/2}\,\phi_k(t)\Big]\right| \\ &\le \frac{1}{4}\left\|V_t^{1/2}\big(\theta-\hat{\theta}(t)\big)\right\|\left\|V_t^{-1/2}\phi_k(t)\right\| = \frac{1}{4}\left\|\hat{\theta}(t)-\theta\right\|_{V_t}\|\phi_k(t)\|_{V_t^{-1}}, \end{aligned} \tag{10}$$

where the first inequality is by Lipschitz property of the logistic function, and the second inequality is by Cauchy-Schwarz inequality $|x^T\,y| \le \|x\|_{M^{-1}}\,\|y\|_M$ for any vectors $x, y$ and matrix $M$. We now bound the last expression in the following two steps.

**Step 1 (Online estimator of parameter $\theta$)**: Recall that $\mathcal{R}_k(t)$ is the stochastic reward feedback of patient $t$ with $\mathcal{R}_k(t) = \sigma\big(\phi_k(t)^\top\theta - f_k(p_k(t);\pi)\big) + \xi_k(t)$, where $\xi_k(t)$ is an independent 1-sub-Gaussian variable.

In the stochastic delay setting, the patient's feedback outcomes are not realized immediately once a decision is made for the patient. Instead, they arrive sequentially with a *stochastic* delay, meaning that there is uncertainty in the timing of when the true feedback will be observed. Consequently, the estimators for the model parameters are updated *on-the-fly* after each patient based on the available information at that time. This dynamic updating ensures that we use the information of a patient to whom a medication with a dose was prescribed previously only if their feedback is realized up to the current time. In other words, the estimator of $\theta$ at each time period $t$ incorporates *only* the pairs of realized information (patient features and feedback outcomes) in the history up to the time period $t$.

We denote $\mathcal{M}(t) = \underset{k\in\mathcal{K}}{\cup}\mathcal{M}_k(t)$ as the set of time-stamps with *realized* feedback outcomes by the end of time period $t-1$, where $\mathcal{M}_k(t) = \{s \mid s \le t-1, s+D(s) \le t-1, k(s)=k\}$ and $k(s)$ is the medication selected by the algorithm at time period $s$. We also denote $\mathcal{N}(t) = \underset{k\in\mathcal{K}}{\cup}\mathcal{N}_k(t)$ as the set of time-stamps with *unrealized* feedback outcomes by the end of time period $t-1$, where $\mathcal{N}_k(t) = \{s \mid s \le t-1, s+D(s) > t-1, k(s)=k\}$.

Let $\hat{\theta}(t)$ be the maximum likelihood estimator (MLE) of $\theta \in \mathbb{R}^{d_1}$ at time period $t$. Given only the realized feedback outcomes at time period $t$, the log-likelihood function $\mathcal{L}_t(\theta)$ is as follows:

$$\mathcal{L}_t(\theta) = \sum_{s\in\mathcal{M}(t)}\Big(\mathcal{R}_k(s)\log\sigma\big(\phi_k(s)^\top\theta - f_k\big(p_k(s);\pi\big)\big) + (1-\mathcal{R}_k(s))\log\big(1-\sigma\big(\phi_k(s)^\top\theta - f_k\big(p_k(s);\pi\big)\big)\big)\Big),$$

where $\mathcal{L}_t(\theta)$ is a strictly concave function of $\theta$.

The MLE $\hat{\theta}(t)$ can then be written as the solution to the following estimating equation:

$$\nabla_\theta \mathcal{L}_t(\theta) = \sum_{s\in\mathcal{M}(t)} \Big(\mathcal{R}_k(s) - \sigma\Big(\phi_k(s)^\top\theta - f_k\big(p_k(s);\pi\big)\Big)\Big)\,\phi_k(s) = 0. \tag{11}$$

Note that this estimator gets updated after each patient by using only the *available* information up to the current time. We next define the *vector-valued error function* $F_t(\theta_0):\mathbb{R}^{d_1}\to\mathbb{R}^{d_1}$ for any parameter $\theta_0$ as:

$$F_t(\theta_0) := \sum_{s\in\mathcal{M}(t)} \Big(\sigma\Big(\phi_k(s)^\top\theta_0 - f_k\big(p_k(s);\pi\big)\Big) - \sigma\Big(\phi_k(s)^\top\theta - f_k\big(p_k(s);\pi\big)\Big)\Big)\,\phi_k(s),$$

which is the difference in the gradients of the log-likelihood function $\mathcal{L}_t(\theta)$ evaluated at any $\theta_0$ and true $\theta$, respectively. Note that $F_t(\theta)=0$ for true parameter $\theta$, and $F_t(\hat{\theta}(t))$ is given by definition as:

$$F_t(\hat{\theta}(t)) = \sum_{s\in\mathcal{M}(t)} \xi_k(s)\,\phi_k(s).$$

Now, the *mean value theorem for vector-valued functions* implies that for any parameter $\theta_1,\theta_2\in\mathbb{R}^{d_1}$, there exists $\bar{\theta}=u\theta_1+(1-u)\theta_2\in\mathbb{R}^{d_1}$ with some $u\in(0,1)$ such that:

$$F_t(\theta_1) - F_t(\theta_2) = \left(\int_0^1 \nabla F_t(\bar{\theta})\,du\right)(\theta_1-\theta_2).$$

Furthermore, let $H_t(\bar{\theta}) = \int_0^1 \nabla F_t(\bar{\theta})\,du = \int_0^1 \sum_{s\in\mathcal{M}(t)} \nabla\sigma\big(\phi_k(s)^\top\bar{\theta} - f_k(p_k(s);\pi)\big)\,\phi_k(s)\,\phi_k(s)^\top\,du$, and also $W_t = \sum_{s\in\mathcal{M}(t)}\phi_k(s)\,\phi_k(s)^\top$ be the design matrix corresponding to *only* the observed features whose feedback outcomes have been received by time $t-1$. Since $\lambda_{\min}(W_t)\geq 1$ and $c_\sigma>0$, we have $H_t(\bar{\theta})\succeq c_\sigma\,W_t\succ 0$. Thus, we have the following for any $\theta_1\neq\theta_2$:

$$(\theta_1-\theta_2)^\top(F_t(\theta_1)-F_t(\theta_2)) = (\theta_1-\theta_2)^\top H_t(\bar{\theta})\,(\theta_1-\theta_2) \geq (\theta_1-\theta_2)^\top c_\sigma\,W_t\,(\theta_1-\theta_2) > 0.$$

The above implies that $F_t(.)$ is an injection from $\mathbb{R}^{d_1}$ to $\mathbb{R}^{d_1}$, and so, $F_t^{-1}$ is well-defined. Thus, (11) has a unique solution $\hat{\theta}(t)=F_t^{-1}(S_t)$ as the MLE for parameter $\theta$, where $S_t := F_t(\hat{\theta}(t)) = \sum_{s\in\mathcal{M}(t)}\xi_k(s)\,\phi_k(s)$.

**Step 2 (High probability confidence bound)**: Using the proposed estimator $\hat{\theta}(t)$, we now derive a high-probability confidence bound for $\left\|\hat{\theta}(t)-\theta\right\|_{V_t}$ in (10). To this aim, following an argument made by Zhou et al. (2019), this term can be decomposed as follows:

$$\begin{aligned}
\left\|\hat{\theta}(t)-\theta\right\|_{V_t}^2 &= \Big(\hat{\theta}(t)-\theta\Big)^\top V_t\,\Big(\hat{\theta}(t)-\theta\Big)\\
&= \Big(\hat{\theta}(t)-\theta\Big)^\top\Big(W_t + \sum_{s\in\mathcal{N}(t)}\phi_k(s)\,\phi_k(s)^\top\Big)\Big(\hat{\theta}(t)-\theta\Big)\\
&\leq \Big(\hat{\theta}(t)-\theta\Big)^\top W_t\,\Big(\hat{\theta}(t)-\theta\Big) + \sum_{s\in\mathcal{N}(t)}\left\|\hat{\theta}(s)-\theta\right\|^2\|\phi_k(s)\|^2\\
&\leq \left\|\hat{\theta}(t)-\theta\right\|_{W_t}^2 + N(t)\,c_\phi^2\left\|\hat{\theta}(t)-\theta\right\|^2,
\end{aligned} \tag{12}$$

where we use the fact that $\|\phi_k(s)\|\leq c_\phi$ and $N(t)=|\mathcal{N}(t)|$.

Note that in (12), $N(t)=\sum_{s=1}^{t-1}\mathbb{1}\{s+D(s)\geq t\}$ is the random number of *unrealized* feedback outcomes at time period $t$. In Lemma 9 (see Appendix C), by building a sequence of stopping times for the number

of realized feedback outcomes, the tail behavior of $N(t)$ can be characterized so that it is sub-Gaussian and upper bounded by $N(t) \le 2\mu_D + \tilde{\sigma}\sqrt{2\log(1/\delta)} + c$ with probability $1-\delta$, where $c = 2\,\tilde{\sigma}^2 \log\left(2\sigma_D^2 + 1\right) + 1$ and $\tilde{\sigma} = \sigma_D\sqrt{p+2}$.

We next develop high probability bounds for expressions $\left\|\hat{\theta}(t) - \theta\right\|_{W_t}^2$ and $\left\|\hat{\theta}(t) - \theta\right\|^2$ in (12), separately.

To bound the first expression, we make the following argument:

$$\begin{aligned}\left\|F_t(\hat{\theta}(t))\right\|_{W_t^{-1}}^2 &= \left\|F_t(\hat{\theta}(t)) - F_t(\theta)\right\|_{W_t^{-1}}^2 \\ &= \left(F_t(\hat{\theta}(t)) - F_t(\theta)\right)^\top W_t^{-1}\left(F_t(\hat{\theta}(t)) - F_t(\theta)\right) \\ &\ge c_\sigma^2\left\|\hat{\theta}(t) - \theta\right\|_{W_t}^2, \end{aligned} \tag{13}$$

where the first equality is established by $F_t(\theta) = 0$, the inequality is established by using the mean value theorem and by $H_t(\bar{\theta}) \succeq c_\sigma W_t$ as well.

Let $S_t = F_t(\hat{\theta}(t))$ for convenience and $\overline{\mathcal{H}}_t$ be a sigma algebra generated by the feature vectors and the noise values upon the arrival of the patient $t$. The following bound then holds with probability at least $1-\delta$ for each time period $t \ge t_0$ and any $\delta > 0$:

$$\left\|\hat{\theta}(t) - \theta\right\|_{W_t}^2 \le \frac{1}{c_\sigma^2}\|S_t\|_{W_t^{-1}}^2 \le \frac{1}{c_\sigma^2}\left(d_1 \log\left(\frac{(t - N(t))\,c_\phi^2}{d_1}\right) + \log\left(\frac{1}{\delta^2}\right)\right), \tag{14}$$

where the first inequality is established by (13), and the second one is obtained by Lemma 1. Indeed, since the noises $\xi_k(s)$ are independent 1-sub-Gaussian random variables, the sequence $\{S_t = \sum_{s\in\mathcal{M}(t)} \xi_k(s)\,\phi_k(s)\}_{t\in\mathcal{T}}$ is a vector-valued *martingale* adapted to $\{\overline{\mathcal{H}}_t\}_{t\in\mathcal{T}}$. Accordingly, Lemma 1 proves that this martingale stays close to zero with high probability (see Lemma 1 in Appendix A).

Moreover, to bound the second expression $\left\|\hat{\theta}(t) - \theta\right\|^2$ in (12), Lemma 2 proves that $\left\|\hat{\theta}(t) - \theta\right\| \le 1$ holds with probability at least $1-\delta$ for any time period $t \ge t_1$, where $t_1$ is the time step for which we have that $\lambda_{\min}(W_{t_1}) \ge \max\{1, 1/c_\sigma^2\left(d_1\log(1 + Tc_\phi^2/d_1) + \log(1/\delta^2)\right)\}$ (see Lemma 2 in Appendix A).

Replacing this bound and (14) in (12) results in the following bound with probability at least $1-2\delta$:

$$\begin{aligned}\left\|\hat{\theta}(t) - \theta\right\|_{V_t} &\le \left\|\hat{\theta}(t) - \theta\right\|_{W_t} + c_\phi\sqrt{N(t)}\left\|\hat{\theta}(t) - \theta\right\|, \\ &\le \frac{1}{c_\sigma}\sqrt{d_1\log\left(\frac{(t-N(t))\,c_\phi^2}{d_1}\right) + \log\left(\frac{1}{\delta^2}\right)} + c_\phi\sqrt{N(t)}.\end{aligned}$$

Putting the above bound in (10), we can upper bound term (I) with probability at least $1-2\delta$.

**Term (II)**: To derive a high probability bound for this term, we first decompose it as follows:

$$\begin{aligned}&\left|\sigma\left(\phi_k(t)^\top\hat{\theta}(t) - f_k\left(p_k(t);\pi\right)\right) - \sigma\left(\phi_k(t)^\top\hat{\theta}(t) - f_k\left(p_k(t);\hat{\pi}(t)\right)\right]\right)\Big| \\ \le\ &\frac{1}{4}\left|\,f_k\left(p_k(t);\pi\right) - f_k\left(p_k(t);\hat{\pi}(t)\right)\right| \le \frac{\max\{\alpha_k,\beta_k\}}{4}\left|\,\chi_k(t)^\top\pi - \chi_k(t)^\top\hat{\pi}(t)\right| \\ \le\ &\frac{\max\{\alpha_k,\beta_k\}}{4}\,\|\chi_k(t)\|_{U_t^{-1}}\,\|\hat{\pi}(t) - \pi\|_{U_t}.\end{aligned} \tag{15}$$

The first and second inequalities are respectively by Lipschitz properties of the logistic function $\sigma(.)$ with constant $1/4$ and the penalty function $f_k(.)$ with constant $\max\{\alpha_k,\beta_k\}$, which can be shown by the reverse triangle inequality $|\,\|x\| - \|y\|\,| \le \|x-y\|$ for $x, y \in \mathbb{R}^n$. The last inequality is by Cauchy-Schwarz inequality.

To bound the expression on the right hand side of (15), we follow the same two-step procedure that we took for bounding term (I) in the previous part as follows.

Recall that $\mathcal{P}_k(t)$ is the sub-outcome with $\mathcal{P}_k(t) = \chi_k(t)^\top \pi + \varepsilon_k(t)$, where the noise $\varepsilon_k(t)$ is an independent $\lambda$-sub-Gaussian variable. We define $\hat{\pi}(t)$ as the maximum-likelihood estimator of parameter $\pi$ at time period $t$ by using *only* the available information (features and sub-outcomes) pairs up to the current time $t$. Since the linear model can be viewed as a special case of the generalized linear model with a *linear* link function, this estimator $\hat{\pi}(t)$ can be obtained by solving the following estimating equation:

$$\nabla_\pi \mathcal{U}_t(\pi) = \sum_{s \in \mathcal{M}(t)} \Big( \mathcal{P}_k(s) - \chi_k(s)^\top \pi \Big) \chi_k(s) = 0,$$

where $\mathcal{U}_t(\pi)$ is the log-likelihood function for estimating $\pi$.

Next, following the same two steps for bounding term (I), we establish the following bound for the MLE $\hat{\pi}(t)$ of the parameter $\pi$. This bound holds with probability at least $1-2\delta$ for any time period $t \geq t_2$, where $t_2$ is the time step for which we have $\lambda_{\min}(Z_{t_2}) \geq \max\big\{1, \lambda^2\big((d_2+K)\log(Tc_\chi^2/(d_2+K)) + \log(1/\delta^2)\big)\big\}$, where $Z_t = \sum_{s \in \mathcal{M}(t)} \chi_k(s)\, \chi_k(s)^\top$ is the design matrix corresponding to only the observed features whose feedback outcomes have been received by time $t-1$:

$$\|\hat{\pi}(t) - \pi\|_{U_t} \leq \lambda \sqrt{(d_2+K)\log\left(\frac{(t-N(t))\, c_\chi^2}{d_2+K}\right) + \log\left(\frac{1}{\delta^2}\right)} + c_\chi \sqrt{N(t)}.$$

Putting the above bound in (15), we can upper bound term (II) with probability at least $1-2\delta$.

Finally, plugging all the above derived results for term (I) as well as the results for term (II) into the decomposition (9) completes the proof where we set $t_0 = \max\{t_1, t_2\}$. **Q.E.D.**

Proposition 5 **(Confidence Bound on Expected Reward under Constant Delays)**. *If the true feedback outcomes are revealed after $\Delta$ time periods for an individual, then for any $t \geq \Delta$ and $\delta > 0$, the following bound on the difference between the true and estimated expected rewards for each selected medication $k$ holds with probability at least $1-2\delta$:*

$$\Big| \mathbf{V}_t^{(k,y)}\big(\theta, \pi\big) - \mathbf{V}_t^{(k,y)}\big(\hat{\theta}(t), \hat{\pi}(t)\big) \Big| \leq \widetilde{\mathrm{Rad}}_k(t),$$

*where $\hat{\theta}(t)$ and $\hat{\pi}(t)$ are respectively the regularized maximum likelihood estimators of the model parameters $\theta$ and $\pi$ at time period $t$ obtained under constant delays, and the confidence radius $\widetilde{\mathrm{Rad}}_k(t)$ is defined as:*

$$\begin{aligned}\widetilde{\mathrm{Rad}}_k(t) = \frac{1}{4\widetilde{c}_\sigma} \|\phi_k(t-\Delta)\|_{V_{t-\Delta}^{-1}} \left( \sqrt{2\log\left(\frac{\det(V_{t-\Delta})^{1/2}\det(\gamma I)^{-1/2}}{\delta}\right)} + \kappa \right) \\ + \frac{\nu}{2} \|\chi_k(t-\Delta)\|_{U_{t-\Delta}^{-1}} \left( \sqrt{2\lambda^2 \log\left(\frac{\det(U_{t-\Delta})^{1/2}\det(\varrho I)^{-1/2}}{\delta}\right)} + \varrho^{1/2} \right),\end{aligned}$$

*where $\widetilde{c}_\sigma = \inf_{\theta, \pi, \phi_k, p_k} \nabla_\theta \sigma\big(\phi_k^\top \theta - f_k(p_k; \pi)\big)$, $V_{t-\Delta} = \sum_{s=1}^{t-\Delta} \phi_k(s)\, \phi_k(s)^\top + \gamma I$ and $U_{t-\Delta} = \sum_{s=1}^{t-\Delta} \chi_k(s)\, \chi_k(s)^\top + \varrho I$ are the design matrices corresponding to the observed features by $t-\Delta$. Also, $\kappa$ and $\zeta$ are regularization parameters for the estimators $\hat{\theta}(t)$ and $\hat{\pi}(t)$, respectively, and $\nu = \max\{\alpha_k, \beta_k\}/2$.*

*Proof of Proposition 5:* We need to establish high probability bounds for both terms (I) and (II) of the same decomposition argument (9) made in the proof of Proposition 4. Consider that $k$ is the medication selected by our algorithm for patient $t$. For now, assume that there is no delay in observing feedback outcomes.

**Term (I)**: We obtain a high probability confidence bound for this term in the following two steps.

**Step 1 (Online estimator of parameter $\theta$)**: We formally define $\hat{\theta}(t)$ as the regularized maximum likelihood estimator of $\theta \in \mathbb{R}^{d_1}$ at time period $t$. Under the *no delayed feedback assumption*, the regularized log-likelihood function of the available data at time period $t$ is derived as follows:

$$\mathcal{L}_t(\theta) = \sum_{s=1}^{t-1} \mathcal{R}_k(s) \log \sigma\left(\phi_k(s)^\top \theta - f_k\left(p_k(t);\pi\right)\right) + (1-\mathcal{R}_k(s)) \log\left(1-\sigma\left(\phi_k(s)^\top \theta - f_k\left(p_k(t);\pi\right)\right)\right) - \frac{\kappa}{2}\|\theta\|^2,$$

where $\kappa$ is the regularization parameter and $\mathcal{L}_t(\theta)$ is strictly concave function of $\theta$ for $\kappa > 0$.

We then need to find the maximum of $\mathcal{L}_t(\theta)$ to obtain the regularized maximum likelihood estimator $\hat{\theta}(t)$. To this aim, we set the gradient of $\mathcal{L}_t(\theta)$ to zero as follows:

$$\nabla_\theta \mathcal{L}_t(\theta) = \sum_{s=1}^{t-1} \left(\mathcal{R}_k(s) - \sigma\left(\phi_k(s)^\top \theta - f_k\left(p_k(t);\pi\right)\right)\right) \phi_k(s) - \kappa\theta = 0.$$

Since $\mathcal{L}_t(\theta)$ is a concave function of $\theta$ for $\kappa > 0$, $\hat{\theta}(t)$ is the unique solution of $\nabla_\theta \mathcal{L}_t(\theta) = 0$.

**Step 2 (High-probability confidence bound)**: Using the proposed estimator $\hat{\theta}(t)$, we establish a high-probability confidence bound for the true expected reward at each time period $t$ by the following steps.

First, we define the design matrix $V_t$ and the vector-valued function $h_t(\theta)$ at each time period $t$ as follows:

$$h_t(\theta) = \sum_{s=1}^{t-1} \sigma\left(\phi_k(s)^\top \theta - f_k\left(p_k(t);\pi\right)\right) \phi_k(s) + \kappa\theta,$$
$$V_t = \sum_{s=1}^{t-1} \phi_k(s)\,\phi_k(s)^\top + \gamma I,$$

where we set $\kappa = \widetilde{c}_\sigma \gamma > 0$. Note that $V_t$ contains the first $t-1$ time-steps of observed features.

Next, according to the mean-value theorem and the Lipschitz property of the logistic function, Lemma 7 (see Appendix C) establishes the following almost-surely bound:

$$\left|\sigma\left(\phi_k(t)^\top \theta - f_k\left(p_k(t);\pi\right)\right) - \sigma\left(\phi_k(t)^\top \hat{\theta}(t) - f_k\left(p_k(t);\pi\right)\right)\right| \le \frac{1}{4\widetilde{c}_\sigma} \|\phi_k(t)\|_{V_t^{-1}} \left\|h_t(\theta) - h_t(\hat{\theta}(t))\right\|_{V_t^{-1}}.$$

We now bound the term $\left\|h_t(\theta) - h_t(\hat{\theta}(t))\right\|_{V_t^{-1}}$. First, recall that $\hat{\theta}(t)$ is the unique solution of $\nabla_\theta \mathcal{L}_t(\theta) = 0$; thus, the following equation holds at each time period $t$:

$$\sum_{s=1}^{t-1} \sigma\left(\phi_k(s)^\top \hat{\theta}(t) - f_k\left(p_k(t);\pi\right)\right) \phi_k(s) + \kappa\,\hat{\theta}(t) = \sum_{s=1}^{t-1} \mathcal{R}_k(s)\,\phi_k(s). \tag{16}$$

Consequently, we expand and bound the term $\left\|h_t(\theta) - h_t(\hat{\theta}(t))\right\|_{V_t^{-1}}$ with probability at least $1-\delta$, for any time period $t$ and $\delta > 0$ as follows:

$$\begin{aligned}
&\left\|h_t(\theta) - h_t(\hat{\theta}(t))\right\|_{V_t^{-1}} \\
&= \left\| \sum_{s=1}^{t-1} \left( \sigma\left(\phi_k(s)^\top \hat{\theta}(t) - f_k\left(p_k(t);\pi\right)\right) - \sigma\left(\phi_k(s)^\top \theta - f_k\left(p_k(t);\pi\right)\right) \right) \phi_k(s) + \kappa\left(\hat{\theta}(t) - \theta\right) \right\|_{V_t^{-1}}
\end{aligned}$$

$$= \left\| \sum_{s=1}^{t-1} \left( \mathcal{R}_k(s) - \sigma\Big(\phi_k(s)^\top \theta - f_k\big(p_k(t);\pi\big)\Big) \right) \phi_k(s) - \kappa\theta \right\|_{V_t^{-1}}$$

$$\leq \left\| \sum_{s=1}^{t-1} \xi_k(s)\, \phi_k(s) \right\|_{V_t^{-1}} + \kappa \left\|\theta\right\|_{V_t^{-1}} \leq \sqrt{2\log\left(\frac{\det(V_t)^{1/2}\det(\gamma I)^{-1/2}}{\delta}\right)} + \kappa, \tag{17}$$

where the first equality is by the definition of function $h_t(.)$, and the second equality is by (16). The first inequality is by knowing that $\xi_k(s) = \mathcal{R}_k(s) - \sigma\big(\phi_k(s)^\top\theta - f_k(p_k(t);\pi)\big)$ and the triangle inequality.

To argue about the last inequality in (17), let $\overline{\mathcal{H}}_t$ be a sigma algebra generated by all the feature vectors and the noise values of the patients who arrived by the end of time period $t-1$. Now, since the noises $\xi_k(s)$ are independent 1-sub-Gaussian random variables, then $\{\sum_{s=1}^{t-1}\xi_k(s)\,\phi_k(s)\}_{t\in\mathcal{T}}$ is a vector-valued *martingale* adapted to $\{\overline{\mathcal{H}}_t\}_{t\in\mathcal{T}}$. Following Theorem 1 of Abbasi-Yadkori et al. (2011), this martingale can be bounded with a high probability, that is, it stays close to zero with a high probability, which results in the last inequality in (17). Also, using the inequality $\lambda_{\min}(M)\|x\|_2^2 \leq \|x\|_M^2 \leq \lambda_{\max}(M)\|x\|_2^2$ for any positive definite matrix $M$ and any vector $x \neq 0$ as well as $\lambda_{\max}(M) = \lambda_{\min}^{-1}(M)$, we note that $\|\theta\|_{V_t^{-1}}^2 \leq \lambda_{\min}^{-1}(V_t)\,\|\theta\|^2 \leq 1$ since $\|\theta\| \leq 1$ and $\lambda_{\min}(V_t) \geq \gamma \geq 1$.

Therefore, we establish the following confidence bound on the difference between the true and estimated rewards under estimated $\hat{\theta}(t)$ with probability at least $1-\delta$ for any time period $t$:

$$\left|\sigma\Big(\phi_k(t)^\top\theta - f_k\big(p_k(t);\pi\big)\Big) - \sigma\Big(\phi_k(t)^\top\hat{\theta}(t) - f_k\big(p_k(t);\pi\big)\Big)\right|$$

$$\leq \frac{1}{4\widetilde{c}_\sigma}\left\|\phi_k(t)\right\|_{V_t^{-1}} \left(\sqrt{2\log\left(\frac{\det(V_t)^{1/2}\det(\gamma I)^{-1/2}}{\delta}\right)} + \kappa\right). \tag{18}$$

**Term (II)**: To derive a high probability bound for this term, we first decompose it as follows:

$$\left|\sigma\Big(\phi_k(t)^\top\hat{\theta}(t) - f_k\big(p_k(t);\pi\big)\Big) - \sigma\Big(\phi_k(t)^\top\hat{\theta}(t) - f_k\big(p_k(t);\hat{\pi}(t)\big)\Big)\right|$$

$$\leq \frac{1}{4}\left|f_k\big(p_k(t);\pi\big) - f_k\big(p_k(t);\hat{\pi}(t)\big)\right| \leq \frac{\max\{\alpha_k,\beta_k\}}{4}\left|\chi_k(t)^\top\pi - \chi_k(t)^\top\hat{\pi}(t)\right|, \tag{19}$$

where the first inequality is by the Lipschitz property of logistic function with constant $1/4$, and the second one is by the the Lipschitz properties of dose penalty function with constant $\max\{\alpha_k,\beta_k\}$ (due to the reverse triangle inequality $|\,\|x\| - \|y\|\,| \leq \|x-y\|$ for $x, y \in \mathbb{R}^n$).

We next follow the same two-step procedure that we took for bounding term (I) to derive a high-probability bound for the expression on the RHS of (19).

**Step 1 (Online estimator of parameter $\pi$)**: Under the *no delayed feedback* assumption, we define $\hat{\pi}(t)$ as the regularized least square estimator for parameter $\pi$. According to the least square principle, we derive this estimator by minimizing the following least square expression:

$$\mathcal{U}_t(\pi) = \Big(\sum_{s=1}^{t-1}\chi_k(s)^\top\pi - \mathcal{P}_k(s)\Big)^\top\Big(\sum_{s=1}^{t-1}\chi_k(s)^\top\pi - \mathcal{P}_k(s)\Big) + \varrho\left\|\pi\right\|^2,$$

where $\varrho > 0$ is the regularization parameter. Minimizing the above term by setting $\nabla_\pi\mathcal{U}_t(\pi) = 0$ results in the following estimator for the parameter $\pi$:

$$\hat{\pi}(t) = \Big(\sum_{s=1}^{t-1}\chi_k(s)\,\chi_k(s)^\top + \varrho I\Big)^{-1}\Big(\sum_{s=1}^{t-1}\mathcal{P}_k(s)\,\chi_k(s)\Big).$$

**Step 2 (High probability confidence bound)**: Following the approach used in the proof of Theorem 2 of Abbasi-Yadkori et al. (2011), we can obtain the following bound on the difference between the true and estimated dose effectiveness measures holds:

$$\left|\chi_k(t)^\top \pi - \chi_k(t)^\top \hat{\pi}(t)\right| \leq \|\chi_k(t)\|_{U_t^{-1}} \left( \left\| \sum_{s=1}^{t-1} \varepsilon_k(s)\, \chi_k(s) \right\|_{U_t^{-1}} + \varrho^{1/2} \right), \tag{20}$$

where $U_t = \sum_{s=1}^{t-1} \chi_k(s)\ \chi_k(s)^\top + \varrho I$ is the design matrix corresponding the first $t-1$ time-steps of the observed features.

Now, consider $\overline{\mathcal{H}}_t$ as a sigma algebra generated by the feature vectors and the noise values of the patients who arrived by the end of time period $t-1$. Since $\varepsilon_k(s)$ is an independent $\lambda$-sub-Gaussian variable, then $\{\sum_{s=1}^{t-1} \varepsilon_k(s)\, \chi_k(s)\}_{t\in\mathcal{T}}$ is a *martingale* adapted to $\{\overline{\mathcal{H}}_t\}_{t\in\mathcal{T}}$. Accordingly, with probability at least $1-\delta$, the following bound holds for any time period $t$ and $\delta > 0$ (see Theorem 1 in Abbasi-Yadkori et al. 2011):

$$\left\| \sum_{s=1}^{t-1} \varepsilon_k(s)\, \chi_k(s) \right\|_{U_t^{-1}} \leq \sqrt{2\lambda^2 \log\left( \frac{\det(U_t)^{1/2} \det(\varrho I)^{-1/2}}{\delta} \right)}. \tag{21}$$

Integrating the bounds (20) and (21) with (19), we establish the following bound on the difference between the true and estimated rewards under estimator $\hat{\pi}(t)$ with probability at least $1-\delta$:

$$\begin{aligned} &\left|\sigma\Big(\phi_k(t)^\top \hat{\theta}(t) - f_k\big(p_k(t);\pi\big)\Big) - \sigma\Big(\phi_k(t)^\top \hat{\theta}(t) - f_k\big(p_k(t);\hat{\pi}(t)\big)\Big)\right| \\ &\leq \frac{\max\{\alpha_k, \beta_k\}}{4}\ \|\chi_k(t)\|_{U_t^{-1}} \left( \sqrt{2\lambda^2 \log\left( \frac{\det(U_t)^{1/2} \det(\varrho I)^{-1/2}}{\delta} \right)} + \varrho^{1/2} \right) \end{aligned} \tag{22}$$

Finally, replacing the bounds (18) and (22) in the decomposition argument (9) and using a union bound argument, we obtain a bound on $\left|\mathbf{V}_t^{(k,y)}(\theta,\pi) - \mathbf{V}_t^{(k,y)}\big(\hat{\theta}(t),\hat{\pi}(t)\big)\right|$ with probability at least $1-2\delta$.

However, the resulting bound is under the assumption that the there is no delay in observing feedback outcomes. In §2, under constant delays, we assume that the true feedback outcomes of a patient are realized after $\Delta$ time periods. To handle such delay, we wait for the fixed amount of time $\Delta$, which guarantees that the patient's feedback outcomes are realized, and then update the model parameters. That is, we obtain estimators $\hat{\theta}(t)$ and $\hat{\pi}(t)$ using only the realized information at time period $t-\Delta$. Using this observation, we can derive the final bound. **Q.E.D.**

LEMMA 1 **(Self-Normalized Bound for Vector-Valued Martingale $S_t$).** *Consider that we have the vector sequence $S_t = \sum_{s\in\mathcal{M}(t)} \xi_k(s)\, \phi_k(s)$, where $\phi_k(s) \in \mathbb{R}^{d_1}$, $\|\phi_k(s)\| \leq c_\phi$, and the noise $\xi_k(s)$ is a conditionally $\sigma^2$-sub-Gaussian random variable. Moreover, let $W_t = \sum_{s\in\mathcal{M}(t)} \phi_k(s)\ \phi_k(s)^\top$ be the design matrix corresponding to only the observed features whose feedback outcomes have been received by time $t-1$. Then, the following bound holds with probability at least $1-\delta$:*

$$\|S_t\|^2_{W_t^{-1}} \leq \sigma^2 \left( d_1 \log\left( \frac{(t-N(t))\, c_\phi^2}{d_1} \right) + \log\left( \frac{1}{\delta^2} \right) \right),$$

*for any time period $t \geq t_0$, where $t_0$ is any time period such that $\lambda_{\min}(W_{t_0}) \geq 1$, and $N(t)$ is the number of unrealized feedback outcomes by time $t-1$.*

*Proof of Lemma 1:* We first need to derive an extension of Theorem 1 in Abbasi-Yadkori et al. (2011), which is about the concentration of a certain vector-valued martingale.

Recall that $S_t = \sum_{s\in\mathcal{M}(t)} \xi_k(s)\,\phi_k(s)$ and $W_t = \sum_{s\in\mathcal{M}(t)} \phi_k(s)\,\phi_k(s)^\top$. Let $t_0 = \min\{\tau \geq 1 : W_\tau \succeq W\}$ for some $W \succ 0$. Then, for any $\delta \in (0,1)$ and a stopping time $t \geq 1$ such that $t \geq t_0$ holds almost surely, we have the following, which holds with probability at least $1-\delta$:

$$\|S_t\|^2_{W_t^{-1}} \leq 2\sigma^2 \log\left(\frac{\det(W_t)^{1/2}\det(W_{t_0})^{-1/2}}{\delta}\right).$$

The expression on the right-hand side can be further simplified by the following algebra:

$$\begin{aligned} 2\log\left(\frac{\det(W_t)^{1/2}\,\det(W_{t_0})^{-1/2}}{\delta}\right) &= 2\log\left(\det(W_t)^{1/2}\right) + 2\log\left(\frac{\det(W_{t_0})^{-1/2}}{\delta}\right) \\ &\leq \log\left(\det(W_t)\right) + 2\log\left(\frac{1}{\delta}\right), \end{aligned} \tag{23}$$

where the inequality is by the choice of $t_0$ for which $\lambda_{\min}(W_{t_0}) \geq 1$, so $\det(W_{t_0})^{-1/2} \leq 1$.

Note that since $W_t$ is a positive definite matrix for any $t \geq t_0$, $\det(W_t)$ is equal to the product of its eigenvalues and $\text{trace}\,(W_t)$ is equal to the sum of its eigenvalues. Now, by using the inequality of arithmetic and geometric means i.e., $\frac{1}{n}\sum_{i=1}^{n} x_i \geq \sqrt[n]{\prod_{i=1}^{n} x_i}$, and $\|\phi_k(s)\| \leq c_\phi$, we have the following:

$$\begin{aligned} \det(W_t) = \prod_{i=1}^{d_1}\lambda_i(W_t) &\leq \left(\frac{1}{d_1}\sum_{i=1}^{d_1}\lambda_i(W_t)\right)^{d_1} \\ &= \left(\frac{1}{d_1}\,\text{trace}\,(W_t)\right)^{d_1} \\ &= \left(\frac{1}{d_1}\sum_{s\in\mathcal{M}(t)}\|\phi_k(s)\|^2\right)^{d_1} \leq \left(\frac{(t-N(t))\,c_\phi^2}{d_1}\right)^{d_1}, \end{aligned}$$

where $\lambda_i(W_t)$ is the $i^{\text{th}}$ eigenvalue of the matrix $W_t$.

Replacing the above bound on $\det(W_t)$ in (23) results in the self-normalized bound on the deviation of the martingale $\|S_t\|^2_{W_t^{-1}}$. **Q.E.D.**

LEMMA 2 **(Regularity Conditions for Design Matrices).** *Let $t_1$ be any time period such that:*

$$\lambda_{\min}(W_{t_1}) \geq \max\left\{1, \frac{1}{c_\sigma^2}\Big(d_1\log(Tc_\phi^2/d_1) + \log(1/\delta^2)\Big)\right\}, \tag{24}$$

*where $W_t = \sum_{s\in\mathcal{M}(t)}\phi_k(s)\,\phi_k(s)^\top$ for any $t$ represents the design matrix corresponding to only the observed features $\phi_k(s)$ whose feedback outcomes have been received by time $t-1$. Then, $\left\|\hat{\theta}(t)-\theta\right\| \leq 1$ with probability at least $1-\delta$ for any $t \geq t_1$, where $\hat{\theta}(t)$ is the MLE for true parameter $\theta$ and is obtained by solving (11).*

*Moreover, let $t_2$ be any time period such that:*

$$\lambda_{\min}(Z_{t_2}) \geq \max\left\{1, \lambda^2\Big((d_2+K)\log(Tc_\chi^2/(d_2+K)) + \log(1/\delta^2)\Big)\right\}, \tag{25}$$

*where $Z_t = \sum_{s\in\mathcal{M}(t)}\chi_k(s)\,\chi_k(s)^\top$ for any $t$ is the design matrix corresponding to only the observed features $\chi_k(s)$ whose feedback outcomes have been received by time $t-1$. Then, $\|\hat{\pi}(t)-\pi\| \leq 1$ with probability at least $1-\delta$ for any time period $t \geq t_2$, where $\hat{\pi}(t)$ is the MLE for true parameter $\pi$.*

REMARK 2. In the SGD-MAB algorithm with stochastic delay case, a warm-up period of length $t_0$ is necessary. We may set the value of $t_0$ by following Proposition 1 of Li et al. (2017). However, in our setting, we have two unknown model parameters $\theta$ and $\pi$, which need to be estimated. Our design matrices, corresponding to only the observed features whose feedback outcomes have been received, should possess certain desirable properties for reliable estimation of these model parameters. To address these issues, we set $t_0 = \max\{t_1, t_2\}$ such that $t_1 = c_1\left(d_1 \log(Tc_\phi^2/d_1) + \log(1/\delta)\right)$ and $t_2 = c_2\left((d_2+K)\log(Tc_\chi^2/(d_2+K)) + \log(1/\delta)\right)$, where $c_1$ and $c_2$ are universal constants. By setting $t_0$ in this manner, we ensure that the length $t_0$ of warm-up period is bounded by a *logarithmic* expression in $T$. This allows sufficient time for accurate estimation of the model parameters and ensures that the design matrices meet the desired properties for effective learning within the given time horizon.

*Proof of Lemma 2:* We shall prove that the estimator $\hat{\theta}(t)$ falls into the $\eta$-neighborhood $\mathcal{B}_\eta(\theta)$ of the *true unknown model parameter* $\theta$ with respect to $\ell_2$-norm for some $\eta > 0$ that will be set later.

To this aim, first let $\mathcal{B}_\eta(\theta) = \{\theta_0 : \|\theta_0 - \theta\| \le \eta\}$ be the $\eta$-neighborhood of the true model parameter $\theta$, and also $\partial\mathcal{B}_\eta(\theta) = \{\theta_0 : \|\theta_0 - \theta\| = \eta\}$. We further define:

$$c_{\sigma,\eta} = \inf\left\{\nabla_{\theta_0}\sigma\left(\phi_k^\top\theta_0 - f_k(p_k;\pi_0)\right) : \theta_0 \in \mathcal{B}_\eta(\theta), \|\pi_0\| \le 1, \|\phi_k\| \le c_\phi, \|\chi_k\| \le c_\chi\right\} > 0.$$

From the proof of Proposition 4, recall the *vector-valued error function* $F_t(\theta_0)$ for any parameter $\theta_0$ as:

$$F_t(\theta_0) := \sum_{s\in\mathcal{M}(t)}\left(\sigma\left(\phi_k(s)^\top\theta_0 - f_k\big(p_k(s);\pi\big)\right) - \sigma\left(\phi_k(s)^\top\theta - f_k\big(p_k(s);\pi\big)\right)\right)\phi_k(s),$$

and we had that $F_t(\theta) = 0$ for the true parameter $\theta$ by definition.

Accordingly, the mean value theorem for vector-valued functions implies for any parameter $\theta_0 \in \partial\mathcal{B}_\eta(\theta)$, there is some $\bar{\theta} = u\theta_0 + (1-u)\theta \in \mathcal{B}_\eta(\theta)$ with some $u \in (0,1)$ such that $F_t(\theta_0) - F_t(\theta) = H_t(\bar{\theta})\,(\theta_0 - \theta)$, where $H_t(\bar{\theta}) = \int_0^1 \nabla F_t(\bar{\theta})\,du = \int_0^1 \sum_{s\in\mathcal{M}(t)} \nabla\sigma\big(\phi_k(s)^\top\bar{\theta} - f_k(p_k(s);\pi)\big)\phi_k(s)\phi_k(s)^\top du \succeq c_{\sigma,\eta}W_t \succ 0$ since $c_{\sigma,\eta} > 0$ and $\lambda_{\min}(W_t) \ge 1$. Thus, we can argue that the followings hold for *every* parameter $\theta_0 \in \partial\mathcal{B}_\eta(\theta)$:

$$\begin{aligned}\|F_t(\theta_0)\|^2_{W_t^{-1}} = \|F_t(\theta_0) - F_t(\theta)\|^2_{W_t^{-1}} &= (\theta_0-\theta)^\top H_t(\bar{\theta})\,W_t^{-1}H_t(\bar{\theta})\,(\theta_0-\theta)\\ &\ge c_{\sigma,\eta}^2\,\lambda_{\min}(W_t)\,\|\theta_0-\theta\|^2 = c_{\sigma,\eta}^2\eta^2\lambda_{\min}(W_t). \end{aligned} \tag{26}$$

where the last inequality is by $H_t(\bar{\theta}) \succeq c_{\sigma,\eta}\,W_t$. This implies that $\inf_{\theta_0\in\partial\mathcal{B}_\eta(\theta)}\|F_t(\theta_0)\|_{W_t^{-1}} \ge c_{\sigma,\eta}\eta\sqrt{\lambda_{\min}(W_t)}$. Given that this condition holds, we use Part (b) of the *inverse function lemma* (see Lemma 8 in Appendix C) to establish the following:

$$\left\{\theta_0 \in \mathbb{R}^{d_1} : \|F_t(\theta_0)\|_{W_t^{-1}} \le c_{\sigma,\eta}\eta\sqrt{\lambda_{\min}(W_t)}\right\} \subseteq \mathcal{B}_\eta(\theta) = \{\theta_0 : \|\theta_0-\theta\| \le \eta\}. \tag{27}$$

The above argument implies that for any $\theta_0$, if $F_t(\theta_0)$ and $F_t(\theta)$ are close, then $\theta_0$ and $\theta$ are also close.

On the other hand, using the fact that $S_t = F_t(\hat{\theta}(t)) = \sum_{s\in\mathcal{M}(t)}\xi_k(s)\,\phi_k(s)$, Lemma 1 develops the following bound that holds with probability at least $1-\delta$ for any time period $t \ge t_1$:

$$\left\|F_t(\hat{\theta}(t))\right\|_{W_t^{-1}} \le \sqrt{\lambda^2\left(d_1\log\left(\frac{Tc_\phi^2}{d_1}\right) + \log\left(\frac{1}{\delta^2}\right)\right)}. \tag{28}$$

The above high-probability bound on $\left\|F_t(\hat{\theta}(t))\right\|_{W_t^{-1}}$ shows that $F_t(\hat{\theta}(t))$ and $F_t(\theta)$ are indeed close.

Finally, we need to find a proper $\eta$ such that $\left\|F_t(\hat{\theta}(t))\right\|_{W_t^{-1}} \leq c_{\sigma,\eta}\eta\sqrt{\lambda_{\min}(W_t)}$ to guarantee $\hat{\theta}(t) \in \mathcal{B}_\eta(\theta)$, which means that $\hat{\theta}(t)$ falls into the $\eta$-neighborhood of the true parameter $\theta$ with respect to $\ell_2$-norm. To do so, if we combine (27) with (28), setting $\eta \geq \frac{\sqrt{\lambda^2\left(d_1\log\left(Tc_\phi^2/d_1\right)+\log(1/\delta^2)\right)}}{c_{\sigma,\eta}\sqrt{\lambda_{\min}(W_t)}}$ implies $\left\|\hat{\theta}(t)-\theta\right\| \leq \eta$ with probability at least $1-\delta$ (note that in our problem $\lambda = \frac{1}{2}$). Moreover, since $c_\sigma = c_{\sigma,1}$, we have $c_{\sigma,\eta} \geq c_\sigma$ when $\eta \leq 1$. Therefore, the following bound holds for any time period $t \geq t_1$ with probability at least $1-\delta$:

$$\left\|\hat{\theta}(t)-\theta\right\| \leq \frac{\sqrt{\lambda^2\left(d_1\log\left(Tc_\phi^2/d_1\right)+\log(1/\delta^2)\right)}}{c_\sigma\sqrt{\lambda_{\min}(W_t)}} \leq 1,$$

whenever $\lambda_{\min}(W_t) \geq 1/c_\sigma^2\left(d_1\log(Tc_\phi^2/d_1)+\log(1/\delta^2)\right)$, which is the case in our problem setting, because we have the condition that $\lambda_{\min}(W_t) \geq \lambda_{\min}(W_{t_1})$. This completes the proof for part one of the lemma.

Note that the proof for the second part of this lemma follows the same steps as above since the linear regression is a special case of generalized linear models. **Q.E.D.**

LEMMA 3 **(Bound on Difference between Upper and Lower Bounds of Expected Reward).** *For any $t \geq t_0$ and $\delta > 0$, the following bound holds with probability at least $1-4\delta$:*

$$\sum_{t=t_0+1}^{T}\mathbb{E}\left[\left(UB_k(t)-LB_k(t)\right)\right] \leq \frac{1}{2}\sqrt{2Td_1\log\left(\frac{Tc_\phi^2}{d_1}\right)}\left(\frac{1}{c_\sigma}\sqrt{d_1\log\left(\frac{Tc_\phi^2}{d_1}\right)+\log\left(\frac{1}{\delta^2}\right)}+c_\phi\sqrt{N_{\max}}\right)$$
$$+\nu\sqrt{2T(d_2+K)\log\left(\frac{Tc_\chi^2}{d_2+K}\right)}\left(\lambda\sqrt{(d_2+K)\log\left(\frac{Tc_\chi^2}{d_2+K}\right)+\log\left(\frac{1}{\delta^2}\right)}+c_\chi\sqrt{N_{\max}}\right),$$

*where $UB_k(t)$ and $LB_k(t)$ are the largest and smallest possible estimated quantities for the expected reward of patient $t$ for whom the medication $k$ with the optimal dose $y^*$ is selected by the proposed SGD-MAB algorithm, respectively.*

*Proof of Lemma 3:* Recall that $UB_k(t)$ and $LB_k(t)$ are defined as the sequence of real-valued functions of the history $\mathcal{H}_t$ and the feature vectors $\phi_k(t)$ and $\chi_k(t)$ of patient $t$ as follows:

$$UB_k(t) = \min\left\{1, \max_{(\theta,\pi)\in\Omega_t}\mathbf{V}_t^{(k,y^*)}(\theta,\pi)\right\},$$
$$LB_k(t) = \max\left\{0, \min_{(\theta,\pi)\in\Omega_t}\mathbf{V}_t^{(k,y^*)}(\theta,\pi)\right\},$$

where $\Omega_t$ is the confidence set that contains true parameters $(\theta,\pi)$ with high probability.

By Proposition 5, we establish the following confidence set $\Omega_t$ for the unknown model parameters $(\theta,\pi)$, which holds with probability at least $1-4\delta$ for any $t \geq t_0$ and $\delta > 0$:

$$\Omega_t = \left\{(\theta,\pi)\;\middle|\;\left|\mathbf{V}_t^{(k,y^*)}(\theta,\pi)-\mathbf{V}_t^{(k,y^*)}\left(\hat{\theta}(t),\hat{\pi}(t)\right)\right| \leq \mathrm{Rad}_k(t)\right\},$$

where $\mathrm{Rad}_k(t)$ is defined as follows:

$$\mathrm{Rad}_k(t) = \frac{1}{4}\left\|\phi_k(t)\right\|_{V_t^{-1}}\left(\frac{1}{c_\sigma}\sqrt{d_1\log\left(\frac{(t-N(t))\,c_\phi^2}{d_1}\right)+\log\left(\frac{1}{\delta^2}\right)}+c_\phi\sqrt{N(t)}\right)$$
$$+\frac{\nu}{2}\left\|\chi_k(t)\right\|_{U_t^{-1}}\left(\lambda\sqrt{(d_2+K)\log\left(\frac{(t-N(t))\,c_\chi^2}{d_2+K}\right)+\log\left(\frac{1}{\delta^2}\right)}+c_\chi\sqrt{N(t)}\right).$$

Next, we define the sequences $\overline{UB}_k(t)$ and $\overline{LB}_k(t)$ as follows:

$$\overline{UB}_k(t) = \mathbf{V}_t^{(k,y^*)}\big(\hat{\theta}(t), \hat{\pi}(t)\big) + \mathrm{Rad}_k(t),$$
$$\overline{LB}_k(t) = \mathbf{V}_t^{(k,y^*)}\big(\hat{\theta}(t), \hat{\pi}(t)\big) - \mathrm{Rad}_k(t).$$

Since $UB_k(t) \le \overline{UB}_k(t)$ and $LB_k(t) \ge \overline{LB}_k(t)$, we have that:

$$\Big(UB_k(t) - LB_k(t)\Big) \le \Big(\overline{UB}_k(t) - \overline{LB}_k(t)\Big) \le 2\mathrm{Rad}_k(t).$$

Now, taking the expectation on both sides of the above inequality and summing over all periods result in the following bound:

$$\sum_{t=t_0+1}^{T} \mathbb{E}\Big[\Big(UB_k(t) - LB_k(t)\Big)\Big] \le 2 \sum_{t=t_0+1}^{T} \mathbb{E}\big[\mathrm{Rad}_k(t)\big].$$

The summation of $\mathbb{E}\big[\mathrm{Rad}_k(t)\big]$ over $T$ time periods on the RHS of the above inequality results in:

$$\begin{aligned}\sum_{t=t_0+1}^{T} \mathbb{E}\big[\mathrm{Rad}_k(t)\big] \le \frac{1}{4} \sum_{t=t_0+1}^{T} \|\phi_k(t)\|_{V_t^{-1}} \left(\frac{1}{c_\sigma}\sqrt{d_1 \log\left(\frac{T\, c_\phi^2}{d_1}\right) + \log\left(\frac{1}{\delta^2}\right)} + c_\phi\sqrt{N_{\max}}\right) \\ + \frac{\upsilon}{2} \sum_{t=t_0+1}^{T} \|\chi_k(t)\|_{U_t^{-1}} \left(\lambda\sqrt{(d_2+K)\log\left(\frac{T\, c_\chi^2}{d_2+K}\right) + \log\left(\frac{1}{\delta^2}\right)} + c_\chi\sqrt{N_{\max}}\right).\end{aligned}$$

Based on Lemma 6 (see Appendix C), we have the following almost-surely bound:

$$\sum_{t=t_0+1}^{T} \|\phi_k(t)\|^2_{V_t^{-1}} \le 2\log\left(\frac{\det(V_{T+1})}{\det(V_{t_0+1})}\right).$$

Note that since $V_{T+1} = \sum_{s=1}^{T} \phi_k(s)\,\phi_k(s)^\top$ and each term $\|\phi(s)\| \le c_\phi$, we then have that $\mathrm{trace}(V_{T+1}) = \sum_{s=1}^{T} \mathrm{trace}(\phi_k(s)\,\phi_k(s)^\top) = \sum_{s=1}^{T} \|\phi_k(s)\|^2 \le T c_\phi^2$. Therefore, using the inequality of arithmetic and geometric means, i.e., $\frac{1}{n}\sum_{i=1}^{n} x_i \ge \big(\prod_{i=1}^{n} x_i\big)^{1/n}$, we have the following:

$$\det(V_{T+1}) = \prod_{i=1}^{d_1} \lambda_i(V_{T+1}) \le \left(\frac{1}{d_1}\sum_{i=1}^{d_1} \lambda_i(V_{T+1})\right)^{d_1} = \left(\frac{\mathrm{trace}\,(V_{T+1})}{d_1}\right)^{d_1} \le \left(\frac{T c_\phi^2}{d_1}\right)^{d_1},$$

where $\lambda_i(V_{T+1})$ is the $i^{\text{th}}$ eigenvalue of the matrix $V_{T+1}$.

Moreover, we have that $\det(V_{t_0+1}) = \prod_{j=1}^{d_1} \lambda_j(V_{t_0+1}) \ge \big(\lambda_{\min}(V_{t_0+1})\big)^{d_1} \ge 1$ since $\lambda_{\min}(V_{t_0+1}) \ge 1$. Using the Cauchy-Schwarz inequality $\sum_{t=1}^{T} \|x_t\|_{M_t^{-1}} \le \sqrt{T}\sqrt{\sum_{t=1}^{T} \|x_t\|^2_{M_t^{-1}}}$ for every vector $x_t$ and matrix $M_t$ and the obtained bounds for $\det(V_{T+1})$ and $\det(V_{t_0+1})$, the summation of the induced matrix norm of the feature vector $\phi_k(t)$ is bounded by:

$$\begin{aligned}\sum_{t=t_0+1}^{T} \|\phi_k(t)\|_{V_t^{-1}} &\le \sqrt{T}\sqrt{\sum_{t=t_0+1}^{T} \|\phi_k(t)\|^2_{V_t^{-1}}} \\ &\le \sqrt{2\,T\log\left(\frac{\det(V_{T+1})}{\det(V_{t_0+1})}\right)} \\ &\le \sqrt{2\,T\,d_1\,\log\left(\frac{T\,c_\phi^2}{d_1}\right)}.\end{aligned}$$

A similar bound can also be derived for $\sum_{t=t_0+1}^{T} \|\chi_k(t)\|_{U_t^{-1}}$, where $\chi_k(t) \in \mathbb{R}^{d_2+K}$:

$$\sum_{t=t_0+1}^{T} \|\chi_k(t)\|_{U_t^{-1}} \leq \sqrt{2\,T\,(d_2+K)\,\log\left(\frac{T\,c_\chi^2}{d_2+K}\right)}.$$

Inserting the above two bounds in the bound for $\sum_{t=t_0+1}^{T} \mathbb{E}\big[\mathrm{Rad}_k(t)\big]$ completes the proof. **Q.E.D.**

LEMMA 4 **(Bound on Difference between Expected Reward and its Upper Bound)**. *For any time period $t \geq t_0$ and $\delta > 0$, the following bound holds with probability at least $1-4\delta$:*

$$\sum_{t=t_0+1}^{T} \mathbb{E}\left[\left(\mathbf{V}_t^{(k^*,y^*)}(\theta,\pi) - UB_{k^*}(t)\right)\right] \leq 4T\delta,$$

*where $UB_{k^*}(t)$ is the largest possible estimated value for the expected reward of patient $t$ for whom the optimal medication $k^*$ with the optimal dose $y^*$ is selected.*

*Proof of Lemma 4:* First, notice that since $\mathbf{V}_t^{(k^*,y^*)}(\theta,\pi) \in (0,1)$ and $UB_{k^*}(t) \geq 0$ for any time period $t$, the following bound holds:

$$\sum_{t=t_0+1}^{T} \left(\mathbf{V}_t^{(k^*,y^*)}(\theta,\pi) - UB_{k^*}(t)\right) \leq \sum_{t=t_0+1}^{T} \mathbb{1}\left(\mathbf{V}_t^{(k^*,y^*)}(\theta,\pi) > UB_{k^*}(t)\right).$$

Taking the expectation on both sides of the above inequality results in the following:

$$\sum_{t=t_0+1}^{T} \mathbb{E}\left[\left(\mathbf{V}_t^{(k^*,y^*)}(\theta,\pi) - UB_{k^*}(t)\right)\right] \leq \sum_{t=t_0+1}^{T} \mathbb{P}\left(\mathbf{V}_t^{(k^*,y^*)}(\theta,\pi) > UB_{k^*}(t)\right).$$

By Proposition 5, we establish the following bound for the unknown model parameters $(\theta,\pi)$, which holds with probability at least $1-4\delta$ for any time period $t \geq t_0$ and $\delta > 0$:

$$\left|\mathbf{V}_t^{(k^*,y^*)}(\theta,\pi) - \mathbf{V}_t^{(k^*,y^*)}\big(\hat{\theta}(t),\hat{\pi}(t)\big)\right| \leq \mathrm{Rad}_{k^*}(t).$$

The above high probability statement is equivalent to the following probability:

$$\mathbb{P}\left(\overline{LB}_{k^*}(t) \leq \mathbf{V}_t^{(k^*,y^*)}(\theta,\pi) \leq \overline{UB}_{k^*}(t)\right) \geq 1-4\delta, \tag{29}$$

where the upper and lower bounds $\overline{UB}_{k^*}(t)$ and $\overline{LB}_{k^*}(t)$ are respectively defined as:

$$\overline{UB}_{k^*}(t) = \mathbf{V}_t^{(k^*,y^*)}\big(\hat{\theta}(t),\hat{\pi}(t)\big) + \mathrm{Rad}_{k^*}(t),$$
$$\overline{LB}_{k^*}(t) = \mathbf{V}_t^{(k^*,y^*)}\big(\hat{\theta}(t),\hat{\pi}(t)\big) - \mathrm{Rad}_{k^*}(t).$$

Note that the above definition implies that $\overline{UB}_{k^*}(t) = \max_{(\theta,\pi)\in\Omega_t} \mathbf{V}_t^{(k^*,y^*)}(\theta,\pi)$.

Recalling that $UB_{k^*}(t) = \min\left\{1, \max_{(\theta,\pi)\in\Omega_t} \mathbf{V}_t^{(k^*,y^*)}(\theta,\pi)\right\}$, we can obtain the following bound, which holds with probability at least $1-4\delta$:

$$\sum_{t=t_0+1}^{T} \mathbb{P}\left(\mathbf{V}_t^{(k^*,y^*)}(\theta,\pi) > UB_{k^*}(t)\right) = \sum_{t=t_0+1}^{T} \mathbb{P}\left(\mathbf{V}_t^{(k^*,y^*)}(\theta,\pi) > \overline{UB}_{k^*}(t)\right) \leq 4T\delta,$$

where the inequality holds due to the probability statement (29).

To show that the above equality holds almost surely, we need to prove that $UB_{k^*}(t) = \overline{UB}_{k^*}(t)$. First, notice that if we have $\max_{(\theta,\pi)\in\Omega_t} \mathbf{V}_t^{(k^*,y^*)}(\theta,\pi) < 1$, it implies that $UB_{k^*}(t) < 1$; thus, we have $UB_{k^*}(t) = \overline{UB}_{k^*}(t)$ for sure. Second, since $\mathbf{V}_t^{(k^*,y^*)}(\theta,\pi) \in (0,1)$, the event $\mathbf{V}_t^{(k^*,y^*)}(\theta,\pi) > UB_{k^*}(t)$ implies that we always have $UB_{k^*}(t) < 1$. Therefore, we can argue that $UB_{k^*}(t) = \overline{UB}_{k^*}(t)$ and the above equality holds, which completes the proof. **Q.E.D.**

## B. Technical Results on Sub-Gaussian SGD and B-SGD Losses

*Proof of Proposition 2:* First, we have the following for any point $z_i \in \mathcal{K}$ by convexity of $g_i(.)$:

$$g_i(z^*) \geq g_i(z_i) + \widehat{\nabla} g_i(z_i)\,(z^* - z_i),\ \ \forall i = 1, ..., n,$$

where $\widehat{\nabla} g_i(z_i) \in \partial g_i(z_i) = \big\{h : g_i(u) \geq g_i(z_i) + h^T(u - z_i), \forall u\big\}$, and $\partial g_i(z_i)$ is the set of all *true* sub-gradients of $g_i(.)$ at $z_i$. Using the above property, we establish the following *regret decomposition*:

$$\begin{aligned}\frac{1}{n}\,\mathbb{E}\Big[\sum_{i=1}^{n}\Big(g_i(z_i) - g_i(z^*)\Big)\Big] &\leq \frac{1}{n}\,\mathbb{E}\Big[\sum_{i=1}^{n}\langle\,\widehat{\nabla} g_i(z_i), z_i - z^*\rangle\Big] \\ &= \underbrace{\frac{1}{n}\,\mathbb{E}\Big[\sum_{i=1}^{n}\langle\,\widetilde{\nabla} g_i, z_i - z^*\rangle\Big]}_{\text{Term (I)}} + \underbrace{\frac{1}{n}\,\mathbb{E}\Big[\sum_{i=1}^{n}\langle\,\widehat{\nabla} g_i(z_i) - \widetilde{\nabla} g_i, z_i - z^*\rangle\Big]}_{\text{Term (II)}},\end{aligned} \tag{30}$$

where $\widetilde{\nabla} g_i$ is a stochastic sub-gradient of $g_i(.)$ at $z_i$. Note that we have $\mathbb{E}\big[\widetilde{\nabla} g_i | \mathcal{F}_{i-1}\big] = \widehat{\nabla} g_i(z_i)$ for a *fixed* $z_i$, where the filtration $\mathcal{F}_{i-1} = \sigma\left(\widetilde{\nabla} g_1, \cdots, \widetilde{\nabla} g_{i-1}\right)$ provides past observations.

We next bound each term on the RHS of (30), below.

**Term (I)**: We show term (I) is bounded by the following:

$$\frac{1}{n}\,\mathbb{E}\Big[\sum_{i=1}^{n}\langle\,\widetilde{\nabla} g_i,\, z_i - z^*\rangle\Big] \leq \frac{3}{2}\,\frac{G\zeta}{\sqrt{n}}.$$

To this aim, we first establish the following argument by using the Pythagorean theorem:

$$\|z_{i+1} - z^*\|^2 = \Big\|\mathbf{Proj}_{\mathcal{K}}\Big(z_i - \eta_i\,\widetilde{\nabla} g_i\Big) - z^*\Big\|^2 \leq \Big\|z_i - \eta_i\,\widetilde{\nabla} g_i - z^*\Big\|^2,$$

where the projection operator is defined by $\mathbf{Proj}_{\mathcal{C}}(x) = \arg\min_{y\in\mathcal{C}}\|x - y\|$ for projecting a point $x \in \mathbb{R}^n$ onto a convex set $\mathcal{C}$. Therefore, we have that:

$$\|z_{i+1} - z^*\|^2 \leq \|z_i - z^*\|^2 + \eta_i^2\,\Big\|\widetilde{\nabla} g_i\Big\|^2 - 2\eta_i\,\langle\,\widetilde{\nabla} g_i,\, z_i - z^*\rangle,$$

which implies that we have the following by reorganizing terms:

$$2\,\langle\,\widetilde{\nabla} g_i,\, z_i - z^*\rangle \leq \frac{\|z_i - z^*\|^2 - \|z_{i+1} - z^*\|^2}{\eta_i} + \eta_i\,\Big\|\widetilde{\nabla} g_i\Big\|^2.$$

Thus, summing over all iterations and taking the average, we establish the following arguments:

$$\begin{aligned}\frac{1}{n}\,\mathbb{E}\Big[\sum_{i=1}^{n}\langle\,\widetilde{\nabla} g_i,\, z_i - z^*\rangle\Big] &\leq \frac{1}{2n}\,\mathbb{E}\left[\sum_{i=1}^{n}\left(\frac{\|z_i - z^*\|^2 - \|z_{i+1} - z^*\|^2}{\eta_i} + \eta_i\,\Big\|\widetilde{\nabla} g_i\Big\|^2\right)\right] \\ &\leq \frac{1}{2n}\,\mathbb{E}\left[\sum_{i=1}^{n}\left(\|z_i - z^*\|^2\left(\frac{1}{\eta_i} - \frac{1}{\eta_{i-1}}\right) + \eta_i\,\Big\|\widetilde{\nabla} g_i\Big\|^2\right)\right] \\ &\leq \frac{1}{2n}\left(G^2\left(\frac{1}{\eta_n}\right) + \zeta^2\sum_{i=1}^{n}\eta_i\right) \\ &\leq \frac{1}{2n}\left(\zeta G\sqrt{n} + 2\sqrt{n}G\zeta\right) = \frac{3}{2}\,\frac{G\zeta}{\sqrt{n}},\end{aligned}$$

where the second inequality is by expanding the telescoping series and the third inequality is by the assumption $\mathbb{E}\Big[\Big\|\widetilde{\nabla} g_i\Big\|^2\Big] \leq \zeta^2$. We also define $\frac{1}{\eta_0} = 0$, and the last inequality follows from using the step size $\eta_i = \frac{G}{\zeta\sqrt{i}}$ and the inequality $\sum_{i=1}^{n}\frac{1}{\sqrt{i}} \leq 2\sqrt{n}$.

**Term (II)**: To bound term (II) in the regret decomposition (30), we first recall the filtration $\mathcal{F}_i$, which is a sigma-field $\mathcal{F}_i = \sigma\left(\widetilde{\nabla} g_1, \cdots, \widetilde{\nabla} g_i\right)$ for each iteration $i$, and then establish the following argument:

$$\mathbb{E}\Big[\langle \widehat{\nabla} g_i(z_i) - \widetilde{\nabla} g_i, z_i - z^* \rangle \big| \mathcal{F}_{i-1}\Big] = \mathbb{E}\Big[\langle \widehat{\nabla} g_i(z_i) - \widetilde{\nabla} g_i, z_i - z^* \rangle \big| z_i\Big] = (z_i - z^*)^\top \mathbb{E}\Big[\big(\widehat{\nabla} g_i(z_i) - \widetilde{\nabla} g_i\big) \big| z_i\Big] = 0,$$

where the first equality is because of the projected SGD step $z_i = \mathbf{Proj}_{\mathcal{K}}\big(z_{i-1} - \eta_{i-1}\widetilde{\nabla} g_{i-1}\big)$, and the last equality is because of $\mathbb{E}\big[\widetilde{\nabla} g_i | \mathcal{F}_{i-1}\big] = \widehat{\nabla} g_i(z_i)$ for a fixed $z_i$ (i.e., conditioned on the filtration $\mathcal{F}_{i-1}$, the mean of the noisy sub-gradient is equal to the true sub-gradient). Accordingly, the above property implies that the sequence $\big\{\langle \widehat{\nabla} g_i(z_i) - \widetilde{\nabla} g_i,\ z_i - z^* \rangle\big\}_{i=1}^n$ is indeed a *martingale difference sequence* with respect to the filtration $\big\{\mathcal{F}_i\big\}_{i=1}^n$ (see Definition 4 in Appendix C). This helps us take care of the dependency (or not being i.i.d random variables) among both the decisions $\{z_i\}_{i=1}^n$ and stochastic sub-gradients $\{\widetilde{\nabla} g_i\}_{i=1}^n$.

Furthermore, since $\widetilde{\nabla} g_i$ is a $\rho$-sub-Gaussian random variable, we conclude that $\langle \widehat{\nabla} g_i(y_i) - \widetilde{\nabla} g_i,\ z_i - z^* \rangle$ is a $G^2\rho$-sub-Gaussian martingale difference sequence. By applying the Azuma-Hoeffding's inequality to this sub-Gaussian martingale difference sequence $\big\{\langle \widehat{\nabla} g_i(z_i) - \widetilde{\nabla} g_i,\ z_i - z^* \rangle\big\}_{i=1}^n$ (see Lemma 10 in Appendix C), we have the following high-probability bound:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \langle \widehat{\nabla} g_i(z_i) - \widetilde{\nabla} g_i,\ z_i - z^* \rangle \geq t\right) \leq \exp\left(-\frac{nt^2}{2G^2\rho}\right),$$

which implies that

$$\frac{1}{n}\,\mathbb{E}\Big[\sum_{i=1}^n \langle \widehat{\nabla} g_i(z_i) - \widetilde{\nabla} g_i,\ z_i - z^* \rangle\Big] \leq \sqrt{\frac{2G^2\rho\log(1/\delta)}{n}}$$

holds with probability at least $1-\delta$.

Putting the bounds we developed for terms (I) and (II) together completes the proof. **Q.E.D.**

LEMMA 5 **(Sub-Gaussian Stochastic Sub-gradient)**. *Assume that the moment generating function of $\widetilde{\pi} = (\widetilde{\omega}^\top, \widetilde{\tau}^\top)^\top$ is well-defined. Then, for the dose penalty function $f_k(.)$ defined in (2), there exists a positive constant $\zeta_k$ such that we have the following:*

1. *The stochastic sub-gradient $\widetilde{\nabla} f_k$ defined in Step (4a) of Algorithm 1 is $\rho$-sub-Gaussian random variable.*
2. *The second moment of the stochastic sub-gradient $\widetilde{\nabla} f_k$ is bounded by $\zeta_k^2$, i.e.,*

$$\mathbb{E}\Big[\widetilde{\nabla} f_k^2\Big] \leq \zeta_k^2 \ \text{ where } \zeta_k = \max(\alpha_k, -\beta_k)\sqrt{\mathbb{E}\big[\widetilde{\tau}\big]_k^2}.$$

*Proof of Lemma 5:* Recall that we consider the following dose penalty function $f_k(.)$ in Algorithm 1, where we drop the time index $t$ for convenience in this proof:

$$f_k\big(\psi_k, y_k; \widetilde{\pi}\big) = \alpha_k\Big(\psi_k^\top \widetilde{\omega} + y_k^\top \widetilde{\tau} - q\Big)^+ + \beta_k\Big(q - \psi_k^\top \widetilde{\omega} - y_k^\top \widetilde{\tau}\Big)^+,$$

where $\widetilde{\pi} = (\widetilde{\omega}^\top, \widetilde{\tau}^\top)^\top$ is the sampled parameter. The sub-gradient of this dose penalty function with respect to the $k^{\text{th}}$ element of $y_k$ is calculated as follows:

$$\widetilde{\nabla} f_k = [\widetilde{\tau}]_k\Big[\alpha_k\,\mathbb{1}\Big(\psi_k^\top \widetilde{\omega} + y_k^\top \widetilde{\tau} > q\Big) - \beta_k\,\mathbb{1}\Big(\psi_k^\top \widetilde{\omega} + y_k^\top \widetilde{\tau} < q\Big)\Big].$$

**Part 1**: We shall prove that the stochastic sub-gradient $\widetilde{\nabla} f_k$ is a $\rho$-sub-Gaussian random variable. Using Taylor's series expansion while taking expectation, we have the following expression:

$$\mathbb{E}\Big[\exp\left(s\,\widetilde{\nabla} f_k\right)\Big] = 1 + s\,\mathbb{E}\Big[|\widetilde{\nabla} f_k|\Big] + \frac{s^2\,\mathbb{E}\Big[|\widetilde{\nabla} f_k|^2\Big]}{2!} + \sum_{m=3}^{\infty} \frac{s^m\,\mathbb{E}\Big[|\widetilde{\nabla} f_k|^m\Big]}{m!},$$

which implies that we have the following argument:

$$\frac{\sum_{m=3}^{\infty} \frac{s^m\,\mathbb{E}\Big[|\widetilde{\nabla} f_k|^m\Big]}{m!}}{s^2} = \frac{\mathbb{E}\Big[\exp\left(s\,\widetilde{\nabla} f_k\right)\Big] - s\,\mathbb{E}\Big[|\widetilde{\nabla} f_k|\Big] - \frac{s^2\,\mathbb{E}\Big[\widetilde{\nabla} f_k^2\Big]}{2} - 1}{s^2},$$

which converges to 0 as $s \to 0$ by using L'Hôspital's rule. This means that we have the following:

$$\frac{\sum_{m=3}^{\infty} \frac{s^m\,\mathbb{E}\Big[|\widetilde{\nabla} f_k|^m\Big]}{m!}}{s^2} = o(s^2).$$

Moreover, we note from the stochastic sub-gradient that $|\widetilde{\nabla} f_k| \leq \max(\alpha_k, -\beta_k)\,[\widetilde{\tau}]_k$. Now consider:

$$\begin{aligned}
\mathbb{E}\Big[\exp\left(s\,\widetilde{\nabla} f_k\right)\Big] &\leq 1 + \sum_{m=2}^{\infty} \frac{s^m\,\mathbb{E}\Big[|\widetilde{\nabla} f_k|^m\Big]}{m!} \\
&= 1 + \frac{s^2\,\mathbb{E}\Big[|\widetilde{\nabla} f_k|^2\Big]}{2!} + \sum_{m=3}^{\infty} \frac{s^m\,\mathbb{E}\Big[|\widetilde{\nabla} f_k|^m\Big]}{m!} \\
&\leq 1 + \frac{s^2\,\left(\max(\alpha_k, -\beta_k)\right)^2}{2!}\,\mathbb{E}\Big[[\widetilde{\tau}]_k^2\Big] + \sum_{m=3}^{\infty} \frac{s^m\,\left(\max(\alpha_k, -\beta_k)\right)^m}{m!}\,\mathbb{E}\Big[[\widetilde{\tau}]_k^m\Big] \\
&\leq 1 + \frac{s^2\,\left(\max(\alpha_k, -\beta_k)\right)^2}{2!}\,\mathbb{E}\Big[[\widetilde{\tau}]_k^2\Big] + o(s^2).
\end{aligned}$$

On the other hand, we know that for any $\rho \geq 0$, we have the following expansion:

$$\exp\left(\frac{s^2\,\rho^2}{2}\right) = 1 + \frac{s^2\,\rho^2}{2} + o(s^2).$$

Therefore, if we choose $\rho$ such that $\rho^2 > \left(\max(\alpha_k, -\beta_k)\right)^2\,\mathbb{E}\Big[[\widetilde{\tau}]_k^2\Big]$, then for all $s > 0$, we have:

$$\mathbb{E}\Big[\exp\left(s\,\widetilde{\nabla} f_k\right)\Big] \leq \exp\left(\frac{s^2\,\rho^2}{2}\right),$$

which shows the stochastic sub-gradient $\widetilde{\nabla} f_k$ is a $\rho$-sub-Gaussian (see Definition 1 in §2). It should be noted that because we assume that the moment generating function of $\widetilde{\pi} = (\widetilde{\omega}^\top, \widetilde{\tau}^\top)^\top$ is well-defined, we have the boundedness for all moments of the random variable $\widetilde{\tau}$.

**Part 2**: We also need to show that the sub-gradient $\widetilde{\nabla} f_k$ has a bounded second moment. Since $|\widetilde{\nabla} f_k| \leq \max(\alpha_k, -\beta_k)\,[\widetilde{\tau}]_k$, then we have $\mathbb{E}\Big[\widetilde{\nabla} f_k^2\Big] \leq \zeta_k^2$ where $\zeta_k = \max(\alpha_k, -\beta_k)\sqrt{\mathbb{E}\Big[[\widetilde{\tau}]_k^2\Big]}$. **Q.E.D.**

PROPOSITION 6 **(High Probability Regret Bound for Sub-Gaussian B-SGD)**. *Let* $\{z_i\}_{i=1}^n \in \mathbb{R}^d$ *be the sequence obtained by the sub-Gaussian B-SGD mechanism with respect to convex and L-Lipchitz functions* $f_i(.)$ *with a domain* $\mathcal{K}$*, i.e.,*

$$z_1 \in \mathcal{K} \;\; \textit{and} \;\; z_{i+1} = \mathbf{Proj}_{\mathcal{K}_\vartheta}\Big(z_i - \eta_i\,\widetilde{g}_i\Big), \;\; \forall i = 1, ..., n-1,$$

*where* $\widetilde{g}_i = \frac{d}{\vartheta}\, f_i(z_i + \vartheta u_i)\, u_i$ *is an approximate stochastic gradient of the function* $f_i(.)$ *at* $z_i$, $u_i$ *is a random unit vector sampled from the Euclidean sphere* $\mathbb{S} = \{u \in \mathbb{R}^d \mid \|u\| = 1\}$, $\vartheta > 0$ *is a perturbation parameter, and* $\eta_i$ *is step size at each iteration.*

*We make the following assumptions:*

1. *Diameter of the set* $\mathcal{K}$ *is bounded by a constant* $G$, *i.e.,* $\sup_{z_1, z_2 \in \mathcal{K}} \|z_1 - z_2\| \leq G$. *The absolute value of the function at any point is bounded by a constant* $C$, *i.e.,* $\sup_{z \in \mathcal{K}} |f_i(z)| \leq C$.
2. *The set* $\mathcal{K}$ *contains the Euclidean ball* $\mathbb{B} = \{z \in \mathbb{R}^d \mid \|z\| \leq 1\}$ *centered at the zero vector.*
3. *The set* $\mathcal{K}_\vartheta$ *is the Minkowski set corresponding to the set* $\mathcal{K}$, *defined by* $\mathcal{K}_\vartheta = \{z \mid \frac{1}{(1-\vartheta)} z \in \mathcal{K}\}$.

*If we choose the step sizes* $\eta_i = \frac{G}{d\, n^{3/4}}$, *the following bound holds with probability at least* $1 - \delta$*:*

$$\sum_{i=1}^{n} \Big( \mathbb{E}\big[f_i(z_i + \vartheta u_i)\big] - f_i(z^*) \Big) \leq \frac{dG}{2} \Big( 1 + C^2 + \frac{8L}{d} \Big)\, n^{3/4} + G\sqrt{2dC \log(1/\delta) n^{3/4}},$$

*where* $z^* = \underset{z \in \mathcal{K}}{\arg\min} \sum_{i=1}^{n} f_i(z)$.

*Proof of Proposition 6:* First, note that Flaxman et al. (2005) proposed the following:

$$\widetilde{g}_i = \frac{d}{\vartheta}\, f_i(z + \vartheta u)\, u$$

as a stochastic gradient estimator of $f_i(.)$ at $z$, where $\vartheta > 0$ is a perturbation parameter, and $u$ is a random unit vector randomly selected from the Euclidean sphere $\mathbb{S}$.

Using Stoke's theorem, it can be shown that $\widetilde{g}_i$ is an unbiased gradient estimator of the $\vartheta$-*smoothed* version $\hat{f}_i(z) = \mathbb{E}_{\mathrm{v} \in \mathbb{B}}\big[f_i(z + \vartheta \mathrm{v})\big]$ of any convex (not necessarily differentiable) function $f_i(.)$, where v is randomly selected from the Euclidean ball $\mathbb{B} = \{\mathrm{v} \in \mathbb{R}^d \mid \|\mathrm{v}\| \leq 1\}$ (see Lemma 2.1 of Flaxman et al. 2005). This implies the following for any $\vartheta > 0$:

$$\mathbb{E}_{u \in \mathbb{S}}\Big[ f_i(z + \vartheta u) u \,\Big|\, z \Big] = \frac{\vartheta}{d}\, \nabla \hat{f}_i(z) = \frac{\vartheta}{d}\, \nabla \mathbb{E}_{\mathrm{v} \in \mathbb{B}}\big[f_i(z + \vartheta \mathrm{v})\big],$$

where $\mathbb{S} = \{u \in \mathbb{R}^d \mid \|u\| = 1\}$ is the Euclidean sphere centered at the zero vector.

Given that the value of the function $f_i(.)$ is bounded by a constant $C$ at any point, we have:

$$\|\widetilde{g}_i\| = \left\| \frac{d}{\vartheta}\, f_i(z + \vartheta u)\, u \right\| \leq \frac{d}{\vartheta}\, C.$$

Using a similar argument as we made in Lemma 5, we can prove that the gradient estimator $\widetilde{g}_i$ is $\rho$-sub-Gaussian and its second moment is bounded by $\mathbb{E}\big[\, \|\widetilde{g}_i\|^2 \,\big] \leq \rho^2$, where $\rho = \frac{d}{\vartheta}\, C$. Also, the $\vartheta$-*smoothed* version $\hat{f}_i(.)$ is a good approximation of $f_i(.)$, because we have the following for $z \in \mathcal{K}$:

$$\begin{aligned} \Big| \hat{f}_i(z) - f_i(z) \Big| &= \Big| \mathbb{E}_{\mathrm{v} \in \mathbb{B}}\Big[ f_i(z + \vartheta \mathrm{v}) \Big] - f_i(z) \Big| \\ &\leq \mathbb{E}_{\mathrm{v} \in \mathbb{B}}\big[\, |f_i(z + \vartheta \mathrm{v}) - f_i(z)| \,\big] \\ &\leq L\vartheta\, \mathbb{E}_{\mathrm{v} \in \mathbb{B}}\big[\, \|\mathrm{v}\| \,\big] \leq L\vartheta, \end{aligned} \tag{31}$$

where the equality is by definition of $\hat{f}_i(.)$, the first inequality is by Jensen's inequality, the second inequality is due to $f_i(.)$ being $L$-Lipschitz, and the last inequality holds because $\mathrm{v} \in \mathbb{B}$.

Moreover, note that we project onto the *shrunken* set $\mathcal{K}_\vartheta = \{z \mid \frac{1}{(1-\vartheta)} z \in \mathcal{K}\}$ instead of the original set $\mathcal{K}$ to avoid moving outside of the set $\mathcal{K}$ when we add the random sampling from the Euclidean sphere $\mathbb{S}$. For these two sets $\mathcal{K}$ and $\mathcal{K}_\vartheta$, we have the following properties:

- **Property 1**: The shrunken set $\mathcal{K}_\vartheta$ is convex for any $0<\vartheta<1$.
- **Property 2**: $\forall z\in\mathcal{K}_\vartheta$, $\mathbb{B}_\vartheta(z)=\{x\mid x=z+\vartheta u\}\subseteq\mathcal{K}$, i.e., all balls of radius $\vartheta$ around points $z\in\mathcal{K}_\vartheta$ are contained in $\mathcal{K}$, because $\mathcal{K}$ is convex and $u\mathbb{B}\subseteq\mathcal{K}$, so $\mathcal{K}_\vartheta+\vartheta u\mathbb{B}\subseteq(1-\vartheta)\mathcal{K}+\vartheta\mathcal{K}=\mathcal{K}$.
- **Property 3**: $\forall z\in\mathcal{K}$, $\exists z_\vartheta\in\mathcal{K}_\vartheta$ such that $\|z_\vartheta-z\|\leq\vartheta G$, where $G$ is the diameter of $\mathcal{K}$.

Now, let $z^*=\arg\min_{z\in\mathcal{K}}\sum_{i=1}^n f_i(z)$, and $y_\vartheta^*=\mathbf{Proj}_{\mathcal{K}_\vartheta}(z^*)$. Denote $\hat{f}_i(z_i)=\mathbb{E}_{\mathrm{v}_i\in\mathbb{B}}\big[f_i(z_i+\vartheta\mathrm{v}_i)\big]$ for shorthand, where $\hat{f}_i(z_i)$ is the $\vartheta$-*smoothed* version of $f_i(z_i)$ at $z_i$. Also, let $x_i=z_i+\vartheta u_i$. We can then bound the regret of the B-SGD as follows:

$$\begin{aligned}\sum_{i=1}^n\Big(\mathbb{E}\big[f_i(x_i)\big]-f_i(z^*)\Big)&\leq\sum_{i=1}^n\Big(\mathbb{E}\big[f_i(x_i)\big]-f_i(z_\vartheta^*)\Big)+\vartheta nLG\\&\leq\sum_{i=1}^n\mathbb{E}\Big[f_i\left(z_i\right)\Big]-\sum_{i=1}^n f_i\left(z_\vartheta^*\right)+2\vartheta nLG\\&\leq\sum_{i=1}^n\mathbb{E}\Big[\hat{f}_i\left(z_i\right)\Big]-\sum_{i=1}^n\hat{f}_i\left(z_\vartheta^*\right)+4\vartheta nLG,\end{aligned}\tag{32}$$

where the first inequality holds by Property 3 and $f_i(.)$ being $L$-Lipschitz, the second inequality is established by $|f_i(x_i)-f_i(z_i)|\leq L\|x_i-z_i\|\leq\vartheta L$. Also, note that since the set $\mathcal{K}$ contains the Euclidean ball, the diameter $G$ of this set is greater than 1 (so $2\vartheta nL\leq 2\vartheta nLG$). The third inequality is by (31).

In (32), $\sum_{i=1}^n\mathbb{E}\big[\hat{f}_i\left(z_i\right)\big]-\sum_{i=1}^n\hat{f}_i\left(z_\vartheta^*\right)$ is indeed equal to the high-probability regret of implementing the sub-Gaussian S-SGD mechanism on the functions $\hat{f}_i(.)$ and over $\mathcal{K}_\vartheta$. In Proposition 3, we develop the following high-probability bound with probability $1-\delta$:

$$\sum_{i=1}^n\mathbb{E}\Big[\hat{f}_i\left(z_i\right)\Big]-\sum_{i=1}^n\hat{f}_i\left(z_\vartheta^*\right)\leq\frac{G^2}{2\eta_n}+\frac{\zeta^2}{2}\sum_{i=1}^n\eta_i+\sqrt{2G^2\rho\log(1/\delta)n},\tag{33}$$

where $\zeta=\rho=\frac{d}{\vartheta}C$.

Plugging the high-probability bound (33) into the regret bound (32), we can establish the following bound, which holds with probability $1-\delta$:

$$\begin{aligned}\sum_{i=1}^n\Big(\mathbb{E}\big[f_i(x_i)\big]-f_i(z^*)\Big)&\leq\sum_{i=1}^n\mathbb{E}\Big[\hat{f}_i\left(z_i\right)\Big]-\sum_{i=1}^n\hat{f}_i\left(z_\vartheta^*\right)+4\vartheta nLG\\&\leq\frac{G^2}{2\eta_n}+\frac{\zeta^2}{2}\sum_{i=1}^n\eta_i+\sqrt{2G^2\rho\log(1/\delta)n}+4\vartheta nLG\\&=\frac{G^2}{2\eta_n}+\frac{1}{2}\left(\frac{dC}{\vartheta}\right)^2\sum_{i=1}^n\eta_i+G\sqrt{2\left(\frac{dC}{\vartheta}\right)\log(1/\delta)n}+4\vartheta nLG\\&=\frac{dG}{2}\,n^{3/4}+\frac{dGC^2}{2}\,n^{3/4}+4LG\,n^{3/4}+G\sqrt{2dC\log(1/\delta)\,n^{3/4}}\\&=\frac{dG}{2}\Big(1+C^2+\frac{8L}{d}\Big)\,n^{3/4}+G\sqrt{2dC\log(1/\delta)\,n^{3/4}},\end{aligned}$$

where we set $\eta_i=\frac{G}{d\,n^{3/4}}$ and $\vartheta=\frac{1}{n^{1/4}}$ in the second inequality, which completes the proof. **Q.E.D.**

*Proof of Theorem 2:* Let $\mathbf{V}_t^{(k,y)}(\theta)=\sigma\big(\phi_k(t)^\top\theta-f_k\big(\psi_k(t),y_k(t)\big)\big)$ denote the expected reward. Note that there is no parameter $\pi$ in the general dose penalty function. Similar to the Bayesian regret decomposition in §4.3, we argue the following decomposition:

$$\begin{aligned}\textsc{BayesReg}(T)&=\mathbb{E}\Big[\sum_{t=1}^{T}\Big(\mathbf{V}_t^{(k^*,y^*)}(\theta)-\mathbf{V}_t^{(k,y)}(\theta)\Big)\Big]\\&=\mathbb{E}\Big[\sum_{t=1}^{T}\Big(\mathbf{V}_t^{(k^*,y^*)}(\theta)-\mathbf{V}_t^{(k,y^*)}(\theta)\Big)\Big]+\mathbb{E}\Big[\sum_{t=1}^{T}\Big(\mathbf{V}_t^{(k,y^*)}(\theta)-\mathbf{V}_t^{(k,y)}(\theta)\Big)\Big].\end{aligned}$$

Below, we bound each term for the B-SGD-MAB algorithm (see Appendix E).

**Part I (Contextual learning loss)**: To bound the contextual learning loss, we follow our techniques in the proof of Proposition 1 under stochastic delays. Accordingly, the following bound for the contextual learning loss holds with probability at least $1-2\delta$:

$$\begin{aligned}&\mathbb{E}\Big[\sum_{t=1}^{T}\Big(\mathbf{V}_t^{(k^*,y^*)}(\theta)-\mathbf{V}_t^{(k,y^*)}(\theta)\Big)\Big]\\&\le t_0+\frac{1}{2}\sqrt{2d_1\log\Big(\frac{Tc_\phi^2}{d_1}\Big)}\Big(\frac{1}{c_\sigma}\sqrt{d_1\log\Big(\frac{Tc_\phi^2}{d_1}\Big)+\log\Big(\frac{1}{\delta^2}\Big)}+c_\phi\sqrt{N_{\max}}\Big)+2T\delta\\&=t_0+\sqrt{T}\left(\mathcal{E}_\delta(T)+2\delta\sqrt{T}\right).\end{aligned}$$

It is worth noting that unlike Proposition 1, there is no unknown parameter $\pi$ in the general convex and lipschitz dose penalty function $f_k(.)$ that we considered in §4.6; thus, we do not need to learn such parameters using an online linear regression that we had in Proposition 1. Accordingly, we only need to include the loss we incurred due to learning the parameter $\theta$ in the above derivation using an online logistic regression.

**Part II (B-SGD sub-optimality loss)**: Recall that in the B-SGD-MAB algorithm, to optimize the dose of each medication, we employ the B-SGD mechanism instead of the SGD mechanism. Accordingly, the B-SGD sub-optimality loss is bounded with probability $1-\delta$ as:

$$\begin{aligned}&\mathbb{E}\Big[\sum_{t=1}^{T}\Big(\mathbf{V}_t^{(k,y^*)}(\theta)-\mathbf{V}_t^{(k,y)}(\theta)\Big)\Big]\\&=\mathbb{E}\Big[\sum_{t=1}^{T}\Big(\sigma\Big(\phi_k(t)^\top\theta-f_k\big(\psi_k(t),y_k^*\big)\Big)-\sigma\Big(\phi_k(t)^\top\theta-f_k\big(\psi_k(t),y_k(t)\big)\Big)\Big)\Big]\\&\le\frac{1}{4}\,\mathbb{E}\Big[\sum_{t=1}^{T}\Big(f_k\big(\psi_k(t),y_k(t)\big)-f_k\big(\psi_k(t),y_k^*\big)\Big)\Big]\le\mathcal{J}_\delta(T)\;T^{3/4},\end{aligned}$$

where $y_k^*=\arg\min_{y_k\in[v_k^{LB},v_k^{UB}]}\sum_{t=1}^{T}f_k\left(\psi_k(t),y_k\right)$, and $\mathcal{J}_\delta(T)=\frac{G}{4}\Big(\frac{1+C^2+8L}{2}+\sqrt{2dC\log(1/\delta)1/T^{3/4}}\Big)$.

Note that in above, the first inequality is by the fact that the logistic function is lipschitz with constant $1/4$. The second inequality is established directly by using the result of Proposition 6 for the general convex and lipschitz dose penalty function $f(.)$, in which the dimension of the dose decision is set to $d=1$.

Finally, putting the above two bounds developed in parts I and II together, we can establish the following bound on the Bayesian regret of the B-SGD-MAB algorithm:

$$\textsc{BayesReg}(T)\le t_0+\mathcal{J}_\delta(T)\,T^{3/4}+\mathcal{E}_\delta(T)\sqrt{T}+2\delta T=\widetilde{\mathcal{O}}\Big(T^{3/4}\Big),$$

which completes the proof where we set $\delta = 1/T$ for the case of stochastic delay.

It is worth noting that the same proof procedure can be followed to derive the Bayesian regret of $\widetilde{\mathcal{O}}\left(T^{3/4}\right)$ for the case of constant delays. **Q.E.D.**

## C. Known Results

In this Appendix, we provide some known results from the related literature. For completeness, we provide the readers with self-contained and more expository versions of their original proofs and results as well.

LEMMA 6 **(Upper-Bound on the Summation of Feature Vectors)**. *Let $\{\phi(t)\}_{t=1}^{\infty}$ be a sequence of feature vectors in $\mathbb{R}^{d_1}$ such that $\|\phi(t)\| \leq c_\phi$ and $V_t = \sum_{s=1}^{t-1} \phi(s)\,\phi(s)^\top$. When $\lambda_{min}(V_t)$ is large enough, i.e., $\lambda_{\min}(V_t) \geq \max\{1, c_\phi^2\}$, then the following holds almost surely (Adopted from Lemma 9 in Dani et al. 2008):*

$$\sum_{t=m+1}^{m+n} \|\phi_k(t)\|^2_{V_t^{-1}} \leq 2\log\left(\frac{\det(V_{m+n+1})}{\det(V_{m+1})}\right)$$

*Proof of Lemma 6:* First, recall that the design matrix $V_{m+n+1} \in \mathbb{R}^{d_1 \times d_1}$ corresponding to the first $m+n$ time-steps of the observed features is defined as:

$$V_{m+n+1} = \sum_{s=1}^{m+n-1} \phi_k(s)\,\phi_k(s)^\top + \phi_k(m+n)\,\phi_k(m+n)^\top = V_{m+n} + \phi_k(m+n)\,\phi_k(m+n)^\top.$$

The determinant of $V_{m+n+1}$ can be obtained then as:

$$\begin{aligned}
\det(V_{m+n+1}) &= \det\left(V_{m+n} + \phi_k(m+n)\,\phi_k(m+n)^\top\right) \\
&= \det\left(V_{m+n}^{1/2}\left(I + V_{m+n}^{-1/2}\,\phi_k(m+n)\,\phi_k(m+n)^\top\,V_{m+n}^{-1/2}\right)V_{m+n}^{1/2}\right) \\
&= \det\left(V_{m+n}\right)\,\det\left(I + \left(V_{m+n}^{-1/2}\,\phi_k(m+n)\right)\left(V_{m+n}^{-1/2}\,\phi_k(m+n)\right)^\top\right) \\
&= \det\left(V_{m+n}\right)\left(1 + \|\phi_k(m+n)\|^2_{V_{m+n}^{-1}}\right) \\
&= \det\left(V_{m+1}\right)\left[\prod_{t=m+1}^{m+n}\left(1 + \|\phi_k(t)\|^2_{V_t^{-1}}\right)\right], \qquad (34)
\end{aligned}$$

where the fourth equality holds because all the eigenvalues of a matrix of the form $(I + xx^\top)$ where $x \in \mathbb{R}^n$ are one except the one eigenvalue, which is $1 + \|x\|^2$. The last equality is obtained by recursion.

Taking the logarithm of (34) from both sides results in the following:

$$\sum_{t=m+1}^{m+n} \log\left(1 + \|\phi_k(t)\|^2_{V_t^{-1}}\right) = \log\left(\frac{\det(V_{m+n+1})}{\det(V_{m+1})}\right).$$

Using the inequality $x \leq 2\log(1+x)$ for any $0 \leq x \leq 1$ along with the above result, we have:

$$\begin{aligned}
\sum_{t=m+1}^{m+n} \min\left\{1, \|\phi_k(t)\|^2_{V_t^{-1}}\right\} &\leq 2\sum_{t=m+1}^{m+n} \log\left(1 + \min\left\{1, \|\phi_k(t)\|^2_{V_t^{-1}}\right\}\right) \\
&\leq 2\sum_{t=m+1}^{m+n} \log\left(1 + \|\phi_k(t)\|^2_{V_t^{-1}}\right) \\
&= 2\log\left(\frac{\det(V_{m+n+1})}{\det(V_{m+1})}\right).
\end{aligned}$$

Note that we have $\|\phi_k(t)\|^2_{V_t^{-1}} \leq \lambda^{-1}_{\min}(V_t)\,\|\phi_k(t)\|^2 \leq \lambda^{-1}_{\min}(V_t)\,c^2_\phi$. When $\lambda_{\min}(V_t)$ is large enough (i.e., $\lambda_{\min}(V_t) \geq \max\{1, c^2_\phi\}$) and knowing $c_\phi \geq 1$, we have $\|\phi_k(t)\|^2_{V_t^{-1}} \leq 1$. Accordingly, we have the following:

$$\sum_{t=m+1}^{m+n} \|\phi_k(t)\|^2_{V_t^{-1}} \leq 2\log\left(\frac{\det(V_{m+n+1})}{\det(V_{m+1})}\right),$$

which completes the proof. **Q.E.D.**

LEMMA 7 **(Initial Confidence Bound on Expected Reward)**. *For any time period $t$, the following upper bound on the difference between the true and estimated expected rewards holds (Adopted from the proof of Proposition 1 in Filippi et al. 2010):*

$$\left|\sigma\Big(\phi_k(t)^\top\theta - f_k\big(p_k(t);\pi\big)\Big) - \sigma\Big(\phi_k(t)^\top\hat{\theta}(t) - f_k\big(p_k(t);\pi\big)\Big)\right| \leq \frac{1}{4\widetilde{c}_\sigma}\,\|\phi_k(t)\|_{V_t^{-1}}\left\|h_t(\theta) - h_t(\hat{\theta}(t))\right\|_{V_t^{-1}}.$$

*where $\hat{\theta}(t)$ is the regularized maximum likelihood estimator of $\theta$ at time $t$, $V_t = \sum_{s=1}^{t-1}\phi_k(s)\,\phi_k(s)^\top + \gamma I$, $h_t(\theta) = \sum_{s=1}^{t-1}\sigma\big(\phi_k(s)^\top\theta - f_k(p_k(s);\pi)\big)\,\phi_k(s) + \kappa\theta$, and $\widetilde{c}_\sigma = \inf_{\theta,\pi,\phi_k,p_k} \nabla_\theta\sigma\big(\phi_k(s)^\top\theta - f_k(p_k(s);\pi)\big)$.*

*Proof of Lemma 7:* Let $\mathcal{S} \subset \mathbb{R}^n$ be an open set and also $\omega_1, \omega_2 \in \mathbb{R}^n$. Consider a vector-valued function $F : \mathcal{S} \to \mathbb{R}^n$, and assume that it is continuously differentiable. According to the *mean-value theorem* for vector-valued functions, there exist $\bar{\omega} = \epsilon\,\omega_1 + (1-\epsilon)\,\omega_2$, where $0 < \epsilon < 1$ such that:

$$F(\omega_2) - F(\omega_1) = \left(\int_0^1 \nabla F(\bar{\omega})\,du\right)(\omega_2 - \omega_1).$$

We use the mean-value theorem for the following continuously differentiable vector-valued function,

$$h_t(\theta) = \sum_{s=1}^{t-1}\sigma\Big(\phi_k(s)^\top\theta - f_k(p_k(s);\pi)\Big)\,\phi_k(s) + \kappa\theta,$$

where $\kappa$ is the regularizatoin parameter. Let $\theta^0 = \epsilon\,\theta + (1-\epsilon)\,\hat{\theta}(t)$ with $0 < \epsilon < 1$. Then, we have the following gradient vector:

$$\nabla h_t(\theta^0) = \sum_{s=1}^{t-1}\nabla_{\theta^0}\sigma\big(\phi_k(s)^\top\theta^0 - f_k(p_k(s);\,\pi)\big)\,\phi_k(s)\,\phi_k(s)^\top + \kappa I.$$

Let $H_t(\theta^0) = \int_0^1 \nabla h_t(\theta^0)\,d\lambda$ and recall that $\widetilde{c}_\sigma = \inf_{\theta,\pi,\phi_k,p_k} \nabla_\theta\sigma\big(\phi_k(s)^\top\theta - f_k(p_k(s);\pi)\big) > 0$. This implies that $H_t(\theta^0) \succeq c_\sigma\,V_t \succ 0$, where $V_t = \sum_{s=1}^{t-1}\phi_k(s)\,\phi_k(s)^\top + \gamma I$ is the design matrix corresponding to the first $t-1$ time-steps of the observed features and $\kappa = \widetilde{c}_\sigma\gamma > 0$. Therefore, the matrix $H_t(\theta^0)$ is a positive definite and non-singular matrix. According to the mean-value theorem, we then have:

$$h_t(\theta) - h_t(\hat{\theta}(t)) = H_t(\theta^0)\cdot\Big(\theta - \hat{\theta}(t)\Big) \Rightarrow \Big(\theta - \hat{\theta}(t)\Big) = H_t^{-1}(\theta^0)\,\Big(h_t(\theta) - h_t(\hat{\theta}(t))\Big).$$

Therefore, we can derive the following bound for each time period $t$:

$$\begin{aligned}
&\left|\sigma\Big(\phi_k(t)^\top\theta - f_k\big(p_k(t);\pi\big)\Big) - \sigma\Big(\phi_k(t)^\top\hat{\theta}(t) - f_k\big(p_k(t);\pi\big)\Big)\right| \\
&\leq \frac{1}{4}\left|\phi_k(t)^\top\theta - \phi_k(t)^\top\hat{\theta}(t)\right| \\
&= \frac{1}{4}\left|\phi_k(t)^\top H_t^{-1}(\theta^0)\,\Big(h_t(\theta) - h_t(\hat{\theta}(t))\Big)\right| \\
&\leq \frac{1}{4}\,\|\phi_k(t)\|_{H_t^{-1}(\theta^0)}\left\|h_t(\theta) - h_t(\hat{\theta}(t))\right\|_{H_t^{-1}(\theta^0)} \\
&\leq \frac{1}{4\,\widetilde{c}_\sigma}\,\|\phi_k(t)\|_{V_t^{-1}}\left\|h_t(\theta) - h_t(\hat{\theta}(t))\right\|_{V_t^{-1}},
\end{aligned}$$

where the first inequality holds by the Lipschitz property of the logistic function with constant 1/4. The equality holds by the mean-value theorem. In the second inequality, we use $|a^T \cdot M \cdot b| \leq \|a\|_M \|b\|_M$ where $\|a\|_M = \sqrt{a^T M a}$. Next, recall that $H_t(\theta^0) \succeq c_\sigma V_t \succ 0 \Rightarrow H_t^{-1}(\theta^0) \preceq \frac{1}{c_\sigma} V_t^{-1}$, which implies that the inequality $\|x\|_{H_t^{-1}(\theta^0)} \leq \frac{1}{\sqrt{c_\sigma}} \|x\|_{V_t^{-1}}$ holds for each vector $x \in \mathbb{R}^d$. This is used in the last inequality above. **Q.E.D.**

LEMMA 8 **(Inverse Function Lemma)**. *Let $F$ be a smooth injective function from $\mathbb{R}^d$ to $\mathbb{R}^d$ with $F(x_0) = y_0$. Moreover, define $\mathcal{B}_\eta(x_0) = \{x \in \mathbb{R}^d : \|x - x_0\| \leq \eta\}$ and $\partial\mathcal{B}_\eta(x_0) = \{x \in \mathbb{R}^d : \|x - x_0\| = \eta\}$. If the condition $\inf_{x \in \partial\mathcal{B}_\eta(x_0)} \|F(x) - y_0\| \geq r$ holds, then we have the following two arguments:*

*(a)* $\mathcal{B}_r(y_0) = \{y \in \mathbb{R}^d : \|y - y_0\| \leq r\} \subseteq F(\mathcal{B}_\eta(x_0))$,

*(b)* $F^{-1}(\mathcal{B}_r(y_0)) = \{x \in \mathbb{R}^d : \|F(x) - y_0\| \leq r\} \subseteq \mathcal{B}_\eta(x_0)$,

*(Adopted from Lemma A in Appendix of Chen et al. 1999).*

LEMMA 9 **(Tail Characterization of $N(t)$)**. *Consider a sequence of i.i.d. non-negative random variables $\{D(t)\}_{t=1}^T$ with mean $\mu_D$ that satisfies the regularity condition (4). Define a random variable $N(t) = \sum_{s=1}^{t-1} \mathbb{1}\{s + D(s) \geq t\}$. We have the following statements (Adopted from Proposition 1 in Zhou et al. 2019):*

*(a) $N(t)$ is sub-Gaussian and $N(t) \leq 2\mu_D + \tilde{\sigma}\sqrt{2\log(1/\delta)} + c$ for each time period $t$, with probability $1 - \delta$.*

*(b) For the maximal quantity $N_{\max} = \max_{1 \leq t \leq T} N(t)$, we have the following high probability bound:*
$N_{\max} \leq 2\mu_D + \tilde{\sigma}\left(\sqrt{2\log T} + \sqrt{2\log(1/\delta) + c'(\tilde{\sigma}\sqrt{2\log T} + 1)}\right) + c$ *with probability* $1 - \delta$,

*where $c = 2\,\tilde{\sigma}^2 \log(2\sigma_D^2 + 1) + 1$, $c' = 2\log(2\sigma_D^2 + 1)$ and $\tilde{\sigma} = \sigma_D\sqrt{p+2}$.*

The following standard results and definitions in this Appendix are stated without any proof, and we refer interested readers to the chapter 2 of Wainwright (2019) for their detailed arguments.

DEFINITION 3 (**Martingale**). A sequence of random variables $\{X_i\}_{i=1}^\infty$ is said to be a *martingale* sequence adapted to some other sequence of random variables $\{Z_i\}_{i=1}^\infty$ if we have the following:

1. $X_i$ is a measurable function of the history $\mathcal{H}_i = \{Z_1, Z_2, \cdots, Z_i\}$ for each $i$ (this informally means that $X_i$ is deterministic given history $\mathcal{H}_i$).
2. $\mathbb{E}[X_i \,|\, \mathcal{H}_{i-1}] = X_{i-1}$, i.e., $X_i$ is centered around $X_{i-1}$.
3. $\mathbb{E}[|X_i|] < \infty$ for each $i$.

DEFINITION 4 (**Martingale Difference Sequence**). Assume $\{X_i\}_{i=1}^\infty$ is a martingale sequence adapted to $\{Z_i\}_{i=1}^\infty$ and define the random variable $D_i = X_i - X_{i-1}$, then $\{D_i\}_{i=1}^\infty$ is called a *martingale difference sequence* adapted to $\{Z_i\}_{i=1}^\infty$ (i.e., $\mathbb{E}[|D_i|] < \infty$ and $\mathbb{E}[D_i \,|\, \mathcal{H}_{i-1}] = 0$ where $\mathcal{H}_i = \{Z_1, Z_2, \cdots, Z_i\}$).

DEFINITION 5 (**Sub-Gaussian Martingale**). $\{D_i\}_{i=1}^\infty$ is said to be a *$\sigma^2$-sub-Gaussian martingale difference sequence* adapted to $\{Z_i\}_{i=1}^\infty$ if the following inequality holds almost surely:

$$\mathbb{E}\Big[\exp(\lambda D_i) \,|\, Z_1, Z_2, \cdots, Z_{i-1}\Big] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right), \text{ for all } \lambda \in \mathbb{R}.$$

LEMMA 10 **(Azuma-Hoeffding for Sub-Gaussian Martingale Difference Sequence)**. *Assume that* $\{D_i\}_{i=1}^{\infty}$ *is a* $\sigma^2$*-sub-Gaussian martingale difference sequence adapted to* $\{Z_i\}_{i=1}^{\infty}$*, such that:*

$$\mathbb{E}\Big[\exp\left(\lambda D_i\right)|Z_{i-1}\Big] \leq \exp\left(\frac{\lambda^2\,\sigma^2}{2}\right), \text{ for all } \lambda \in \mathbb{R}.$$

*Then, the following inequalities hold:*

$$\mathbb{P}\Big(\sum_{i=1}^{n} D_i \geq t\Big) \leq \exp\left(-\frac{t^2}{2n\,\sigma^2}\right), \text{ for all } t \geq 0.$$

$$\mathbb{P}\Big(\sum_{i=1}^{n} D_i \leq -t\Big) \leq \exp\left(-\frac{t^2}{2n\,\sigma^2}\right), \text{ for all } t \geq 0.$$

*Or, simply we have:*

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{n} D_i\Big| \geq t\Big) \leq 2\exp\left(-\frac{t^2}{2n\,\sigma^2}\right), \text{ for all } t \geq 0.$$

*As a corollary of the Azuma-Hoeffding inequality, we have the following bound:*

$$\Big|\sum_{i=1}^{n} D_i\Big| \leq \sigma\sqrt{2n\log\left(\frac{2}{\delta}\right)}, \text{ with probability at least } 1-\delta.$$

## D. Belief Updating with Bayesian Inference

We explain the general idea of online Bayesian logistic and linear regressions. We use them in Step (6) of the proposed SGD-MAB algorithm to adaptively update the belief about the unknown parameters.

Consider a training data set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^d$ is a feature vector and $y_i \in \{-1, +1\}$ (failure, success) is a response variable. Assume that the success/failure probability is a parameterized function $\mathbb{P}(y = \pm 1|x) = \sigma(y \cdot \theta^T x)$ with unknown parameter $\theta \in \mathbb{R}^d$, where the link function is chosen as the logistic function $\sigma(u) = \frac{1}{1+\exp(-u)}$. With the assumption that training labels are independently generated given $\theta$, the likelihood is $\mathbb{P}(\mathcal{D}|\theta) = \prod_{i=1}^{n} \sigma(y_i \cdot \theta^T x_i)$. The estimate of $\theta$ can be found by maximizing the likelihood $\mathbb{P}(\mathcal{D}|\theta)$, or equivalently minimizing the regularized negative log-likelihood under $l_2$ regularization (to avoid over-fitting) with parameter $\kappa > 0$:

$$\min_{\theta} \; -\sum_{i=1}^{n} \log(\sigma(y_i \cdot \theta^T x_i)) + \frac{\kappa}{2}\left\|\theta\right\|.$$

It can be proved that this regularized log-likelihood function is concave in $\theta$ for logistic regression. Consequently, various optimization methods (e.g., Newton's and gradient decent algorithms) can be used for solving it. However, we have a *sequential* setting in our problem. If we want to update our estimator for a set of new realized data at each iteration, we should re-optimize the above problem using all the previous realized data, which is computationally inefficient.

To deal with this hurdle, we adopt a Bayesian approach to perform a recursive update for the estimator with each set of new realized data. Consider a prior $\mathbb{P}(\theta)$ for the parameter $\theta$, we apply the Bayes' theorem to obtain the posterior $\mathbb{P}(\theta|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|\theta)\,\mathbb{P}(\theta)}{\mathbb{P}(\mathcal{D})} \propto \mathbb{P}(\mathcal{D}|\theta)\,\mathbb{P}(\theta)$. Unfortunately, exact Bayesian inference for the above linear classifier is not tractable since the evaluation of the posterior involves a product of sigmoid functions. We can either use Markov Chain Monte Carlo methods (Hastings 1970), or analytic approximations to the posterior (Tierney and Kadane 1986).

We apply the Laplace approximation, which deploys a Gaussian approximation to the posterior. This can be obtained by finding the mode of the posterior distribution and then fitting a Gaussian distribution centered at that mode (see Chapter 4 of Bishop 2006). In particular, define the logarithm of the unnormalized posterior distribution:

$$\Psi(\theta|m,Q,\mathcal{D}) = \log \mathbb{P}(\mathcal{D}|\theta) + \log \mathbb{P}(\theta). \tag{35}$$

Since the logarithm of a Gaussian distribution is a quadratic function, we use a second-order Taylor series to $\Psi$ in (35) around its MAP (maximum a posterior) solution $\hat{\theta} = \arg\max_\theta \Psi(\theta|m,Q,\mathcal{D})$:

$$\Psi(\theta) \approx \Psi(\hat{\theta}) - \frac{1}{2}(\theta-\hat{\theta})^T \mathbf{H} (\theta-\hat{\theta}), \tag{36}$$

where $\mathbf{H}$ is the Hessian of the negative log posterior evaluated at $\hat{\theta}$, i.e., $\mathbf{H} = -\nabla^2\Psi(\theta)|_{\theta=\hat{\theta}}$. By exponentiating both sides of (36), we can observe that the Laplace approximation results in a normal approximation to the posterior i.e., $\mathbb{P}(\theta|\mathcal{D}) \approx \mathcal{N}(\theta|\hat{\theta}, \mathbf{H}^{-1})$.

For Gaussian priors $\mathbb{P}(\theta) = \mathcal{N}(\theta|m,Q)$, we have the following from (35):

$$\Psi(\theta|m,Q,\mathcal{D}) = -\frac{1}{2}(\theta-m)^T Q^{-1} (\theta-m) + \sum_{i=1}^{n} \log\big(\sigma(y_i \cdot \theta^T x_i)\big), \tag{37}$$

and the Hessian $\mathbf{H}$ evaluated at $\hat{\theta}$ is obtained as $\mathbf{H} = Q^{-1} + \sum_{j=1}^{n} \frac{e^{-\langle \hat{\theta}, x_j\rangle}}{\left(1+e^{-\langle \hat{\theta}, x_j\rangle}\right)^2} x_j x_j^T$.

Starting from a Gaussian prior $\mathcal{N}(\theta_\ell|p_\ell^1, (q_\ell^1)^{-1})$ with mean $p_\ell^1$ and variance $(q_\ell^1)^{-1}$ for each $\ell \in \{1,\dots,d\}$, the Laplace approximated posterior is $\mathcal{N}(\theta_\ell|p_\ell^t, (q_\ell^t)^{-1})$ after the $t^{\text{th}}$ iteration. Recently, Wang et al. (2016) proposed an online Bayesian logistic regression algorithm that finds the MAP solution (35) to the posterior after observing a set of new realized data $\mathcal{S}_t = \{(x_j, y_j)\}_{j=1}^m$ at iteration $t$ by solving the following optimization problem (via a one-dimensional bisection search method):

$$\theta_{\max} = \arg\max_\theta \ \frac{1}{2}\sum_{\ell=1}^{d} q_\ell^t \left([\theta]_\ell - p_\ell^t\right)^2 + \sum_{j\in\mathcal{S}_t} \log\left(1+e^{-y_j \langle \theta, x_j\rangle}\right).$$

The updated mean is then $p^{t+1} = \theta_{\max}$ and the inverse variance of each weight $\theta_\ell$ is given by the curvature at the mode as $q_\ell^{t+1} = q_\ell^t + \sum_{j\in\mathcal{S}_t} \frac{e^{-\langle \theta_{\max} x_j\rangle}}{\left(1+e^{-\langle \theta_{\max}, x_j\rangle}\right)^2} \left([x_j]_\ell\right)^2$ for $\forall \ell \in \{1,\dots,d\}$ (see Wang et al. 2016 for details).

To conduct the Bayesian inference for online linear regression, we have the same issue as described for the online logistic regression above. Agrawal and Goyal (2013) proposed a Bayesian inference procedure to update the posterior distribution in online linear regression. In particular, consider $B^t = I_d + \sum_{i=1}^{t-1} x_i x_i^T$ and $u^t = (B^t)^{-1}\left(\sum_{i=1}^{t-1} x_i y_i\right)$, where $x_i \in \mathbb{R}^d$ is a feature vector and $y_i$ is a response variable. Then, if the prior for parameter $\pi$ at time $t$ is given by $\mathcal{N}(u^t, (B^t)^{-1})$, then the posterior distribution for $\pi$ at time $t+1$ is $\mathcal{N}(u^{t+1}, (B^{t+1})^{-1})$, which is derived as follows:

$$\begin{aligned}
\mathbb{P}(\pi|y_t) &\propto \mathbb{P}(y_t|\pi)\,\mathbb{P}(\pi) \\
&\propto \exp\Big\{-\frac{1}{2}\big((y_t - \pi^T x_t)^2 + (\pi-u^t)^T B^t (\pi-u^t)\big)\Big\} \\
&\propto \exp\Big\{-\frac{1}{2}\big(y_t^2 + \pi^T x_t x_t^T \pi + \pi^T B^t \pi - 2\pi^T x_t y_t - 2\pi^T B^t u^t\big)\Big\} \\
&\propto \exp\Big\{-\frac{1}{2}\big(\pi^T B^{t+1} \pi - 2\pi^T B^{t+1} u^{t+1}\big)\Big\} \\
&\propto \exp\Big\{-\frac{1}{2}\big(\pi - u^{t+1}\big)^T B^{t+1} \big(\pi - u^{t+1}\big)\Big\} \\
&\propto \mathcal{N}(u^{t+1}, (B^{t+1})^{-1}).
\end{aligned}$$

## E. B-SGD-MAB Algorithm

Here, we present the B-SGD-MAB algorithm under stochastic delay in observing feedback outcomes. Note that the theoretical performance of this algorithm is provided in Theorem 2.

**Initialization**. Use a warm-up period of length $t_0$ or offline data (see Lemma 2 in Appendix A for setting the parameter $t_0$).

**Parameters**. Let $m_\ell^1$ and $(q_\ell^1)^{-1}$ be the mean and variance of the Gaussian prior distribution for the $\ell$-th element of $\theta$ vector. These parameters can be initialized based on some prior beliefs.

**Main Loop**. We proceed in time periods $\mathcal{T}=\{t_0+1,\ldots,T\}$ with the following steps.

**Step 1 (context information)**. Observe the context information $(\varphi^{\mathcal{R}}(t),\varphi^{\mathcal{P}}(t))$ of patient $t$.

**Step 2 (sampling)**. Sample $[\tilde{\theta}(t)]_\ell$ from the posterior distribution $\mathcal{N}(m_\ell^t,(q_\ell^t)^{-1})$, $\forall \ell\in\{1,\ldots,d_1\}$.

**Step 3 (optimization of the nested decisions)**. Having the random samples $\tilde{\theta}(t)$, choose the medication $k(t)$ with corresponding dose $y_k(t)$ for patient $t$ such that:

$$k(t)=\arg\max_{k\in\mathcal{K}}\left\{\sigma\Big(\phi_k(t)^\top\tilde{\theta}(t)-f_k\big(\psi_k(t),y_k(t)\big)\Big)\right\},$$

where $k(t)$ is the medication selected for patient $t$.

**Step 4 (Bandit SGD mechanism)**. Update and calculate the following:

(4a) Calculate the approximate gradient $\widetilde{g}_k(t)=\frac{u_k(t)}{\vartheta}\,f_k\big(\psi_k(t),y_k(t)\big)$ for the medication selected for patient $t$, where $u_k(t)$ is a unit random number and $0<\vartheta<1$.

(4b) Calculate the next period's dose for the medication selected for patient $t$ by:

$$[y_k(t+1)]_k=\mathbf{Proj}_{\Omega_k}\Big([y_k(t)]_k-\eta_k(t)\,\widetilde{g}_k(t)\Big)+\vartheta\,u_k(t),$$

where $\Omega_k=\{y\mid\frac{1}{(1-\vartheta)}y\in[\upsilon_k^{LB},\upsilon_k^{UB}]\}$, $u_k(t)$ is a unit random number, $0<\vartheta<1$, $\eta_k(t)=\frac{G_k}{T^{3/4}}$ is the step size, and $G_k=\upsilon_k^{UB}-\upsilon_k^{LB}$.

**Step 5 (feedback observation)**. For each medication $k\in\mathcal{K}$, obtain $\mathcal{S}_k(t)$ as the set of time-stamps with *newly realized* feedback outcomes at time period $t$, which is calculated by $\mathcal{S}_k(t)=\mathcal{M}_k(t+1)-\mathcal{M}_k(t)$, where the set $\mathcal{M}_k(t)$ contains the time-stamps with realized feedback outcomes by the end of time period $t-1$ corresponding to each medication $k$.

**Step 6 (belief updating).** Leverage the realized feedback outcomes whose time-stamp is in $\mathcal{S}_k(t)$ for each medication $k$ to update the posterior distribution of the $\theta$ parameter.

(6a) Solve the following optimization problem,

$$\rho_{\max}=\arg\max_{\rho}\ \frac{1}{2}\sum_{\ell=1}^{d_1}q_\ell^t\,([\rho]_\ell-m_\ell^t)^2+\sum_{k\in\mathcal{K}}\sum_{s\in\mathcal{S}_k(t)}\log\left(1+e^{-\varpi(\mathcal{R}_k(s))\,(\rho^\top\phi_k(s)-f_k(\psi_k(s),y_k(s)))}\right),$$

where $\varpi:\{0,1\}\to\{-1,1\}$ is a mapping function such that $\varpi(0)=-1$ and $\varpi(1)=1$.

(6b) Update the mean and variance of the posterior distribution for $\theta$ as follows:

$$m^{t+1}=\rho_{\max},q_\ell^{t+1}=q_\ell^t+\sum_{k\in\mathcal{K}}\sum_{s\in\mathcal{S}_k(t)}\frac{e^{-(\rho_{\max}^\top\phi_k(s)-f_k(\psi_k(s),y_k(s)))}}{(1+e^{-(\rho_{\max}^\top\phi_k(s)-f_k(\psi_k(s),y_k(s)))})^2}\Big([\phi_k(s)]_\ell\Big)^2,\ \forall\,\ell\in\{1,\ldots,d_1\}.$$