A Proof of Proposition 1

Proof. Let $X_{i=1\ldots N}$ be sampled i.i.d. from the uniform distribution over $[0,1]$. Let Y_i be the ith smallest value of the set $\{X_i\}_{i=1\ldots N}$. So that $Y_1 < Y_2 < \ldots < Y_N$[2]. Finally, let Z be chosen uniformly at random from the set $\{Y_i\}_{i=1\ldots k}$. We will use f to refer to the probability density function for a continuous random variable.

The density of the $(k+1)$ order statistic follows a Beta distribution and so has the following form.

$$f_{Y_{k+1}}(y) = \frac{1}{B(k+1, N-k)} y^k (1-y)^{N-k-1} \tag{5}$$

where $B(a,b)$ is the beta function.

We now derive the density of Z.

$$f_{Z|Y_{k+1}=y}(z) = \frac{1(0 \le z < y)}{y} \tag{6}$$

$$f_Z(z) = \int_0^1 f_{Z|Y_{k+1}=y}(z) f_{Y_{k+1}}(y)\,\mathrm{d}y \tag{7}$$

$$= \int_0^1 \frac{1(0 \le z < y)}{y} \frac{1}{B(k+1, N-k)} y^k (1-y)^{N-k-1}\,\mathrm{d}y \tag{8}$$

$$= \int_z^1 \frac{1}{B(k+1, N-k)y} y^k (1-y)^{N-k-1}\,\mathrm{d}y \tag{9}$$

$$= \frac{1}{B(k+1, N-k)} \int_z^1 y^{k-1} (1-y)^{N-k-1}\,\mathrm{d}y \tag{10}$$

$$= \frac{B(k, N-k)}{B(k+1, N-k)} \frac{1}{B(k, N-k)} \int_z^1 y^{k-1} (1-y)^{N-k-1}\,\mathrm{d}y \tag{11}$$

$$= \frac{B(k, N-k)}{B(k+1, N-k)} \frac{1}{B(k, N-k)} (B(k, N-k) - B(z; k, N-k)) \tag{12}$$

$$= \frac{B(k, N-k)}{B(k+1, N-k)} (1 - I_z(k, N-k)) \tag{13}$$

$$= \frac{B(N-k, k)}{B(N-k, k+1)} I_{1-z}(N-k, k) \tag{14}$$

$$\propto I_{1-z}(N-k, k) = 1 - I_z(k, N-k) \tag{15}$$

where $B(x; a, b)$ is the incomplete beta function, i.e., $B(x; a, b) = \int_0^x t^{a-1}(1-t)^{b-1} dt$, and $I_x(a, b)$ is the regularized incomplete beta function, i.e., $I_x(a, b) = \frac{B(x;a,b)}{B(a,b)}$.

We now want to show that our k-of-N procedure results in the same density. Let $\mathcal{M} \sim \mathbb{P}(\mathcal{M})$ and let $V = V^{\pi}_{\mathcal{M}}$, so that V is a real-valued random variable. Let F_V be its cumulative distribution function, and define $F_V^{-1}(y) = \inf_{x \in \mathbb{R}}\{F_V(s) \ge y \text{ for } y \in [0,1]\}$ to be its generalized inverse distribution function. Recall that the k-of-N procedure draws N samples from the distribution F_V and chooses uniformly among the k smallest. This procedure can also be achieved by the inverse sampling transform. Take our uniform $[0,1]$ samples $X_{i=1\ldots n}$, and let $V_i = F_V^{-1}(X_i)$. The set $V_{i=1\ldots n}$ are i.i.d. distributed according to F_V. Furthermore, since F_V^{-1} is non-decreasing, the ordering of V_i and X_i are the same. So, $F_V^{-1}(Z)$ has the same distribution as a random sample from the k smallest values in $\{V_i\}_{i=1\ldots n}$. □

B Proof of Theorem 3

Proof. First, observe that we can trivially extend Theorem 2 to the case of a mixture of k-of-N measures. We simply modify the k-of-N game to begin by having a chance node sample a k-of-N

[2]We can assume $Y_i < Y_{i+1}$ because the event that any two X_i's have the same value has measure zero.

measure, unbeknownst to player 1. Since the algorithm in Theorem 2 involves chance-sampled CFR-BR, we can handle this new game by simply sampling a new k-of-N measure on each iteration. The number of iterations required in such a case is unchanged; the time per iteration increases by the sampling time, which is polynomial in the size of the mixture.

Let $\epsilon > 0$ be given. Consider the μ-robust optimization in Equation 4, and note that for all absolutely continuous measures, the optimization is unchanged if we allow $\mathcal{F}$ to be the set of Lebesgue integrable functions. We will construct an absolutely continuous measure $\hat{\mu}$ with density $g_{\hat{\mu}}$ derived from a mixture of $\hat{m}$ measures, each a k-of-N measure, such that

$$|g_\mu - g_{\hat{\mu}}| = \int_0^1 |g_\mu(x) - g_{\hat{\mu}}(x)|\, dx < \frac{\epsilon}{2\Delta}. \tag{16}$$

Therefore, for any Lebesgue integrable function $f \in \mathcal{F}$,

$$\left| \int_{[0,1]} f\, d\mu - \int_{[0,1]} f\, d\hat{\mu} \right| = \left| \int_0^1 f(x) g_\mu(x)\, dx - \int_0^1 f(x) g_{\hat{\mu}}(x)\, dx \right| \tag{17}$$

$$= \left| \int_0^1 f(x)(g_\mu(x) - g_{\hat{\mu}}(x))\, dx \right| \tag{18}$$

$$\leq \int_0^1 |f(x)|\, |g_\mu(x) - g_{\hat{\mu}}(x))|\ dx \tag{19}$$

$$\leq \Delta \int_0^1 |g_\mu(x) - g_{\hat{\mu}}(x))|\ dx \leq \frac{\epsilon}{2} \tag{20}$$

By Theorem 2 we can find an $\frac{\epsilon}{2}$-approximation of our $\hat{\mu}$-robust policy with high probability in polynomial time (polynomial in $\hat{m}$ and N as well) under our conditions. Since the two objectives never differ by more than $\frac{\epsilon}{2}$ this gives us a high probability ϵ-approximation of a μ-robust policy. Thus, it suffices to show how to construct $g_{\hat{\mu}}$ such that $\hat{m}$ and N are polynomial in $\{\Delta, L, m, \frac{1}{\epsilon}\}$.

In constructing $g_{\hat{\mu}}$, we will first approximate g_μ with a piecewise constant function $\hat{g}$ such that $|\hat{g} - g_\mu| < \frac{\epsilon}{4\Delta}$, and then approximate $\hat{g}$ with $g_{\hat{\mu}}$ using a mixture of k-of-N measures such that $|g_{\hat{\mu}} - \hat{g}| < \frac{\epsilon}{4\Delta}$ Let $a_1, \ldots, a_m$ be the boundaries of the pieces of μ, and let $b_1, \ldots, b_{m'}$ be uniformly spaced along the unit interval with $m' = \lceil \frac{4L\Delta}{\epsilon} \rceil$ pieces. Finally, let the boundaries of the pieces of $\hat{g}$, $x_0, \ldots, x_{\hat{m}}$, be the union of $\{a_i\}$, $\{b_i\}$, and $\{0, 1\}$, so that $x_0 = 0$ and $x_{\hat{m}} = 1$. For any $x \in [x_{i-1}, x_i)$, define $\hat{g}(x) = g_\mu(\frac{x_{i-1}+x_i}{2})$. By construction, the interval $[x_{i-1}, x_i)$ is in the same piece of g_μ and the size of the interval $|x_{i-1} - x_i| \leq \frac{4L\Delta}{\epsilon}$, so by the piecewise Lipschitz continuity of g_μ

$$|\hat{g} - g_\mu| \leq \max_{x \in [0,1]} |\hat{g}(x) - g_\mu(x)| \tag{21}$$

$$\leq L \frac{4L\Delta}{\epsilon} = \frac{4\Delta}{\epsilon}, \tag{22}$$

where $\hat{g}$ has $\hat{m} = m + \lceil \frac{4L\Delta}{\epsilon} \rceil$ pieces.

Notice that $\hat{g}$ is a non-increasing function since g_μ is non-increasing. Therefore, we can write $\hat{g}$ as a weighted sum of decreasing unit step functions,

$$\hat{g}(x) = \sum_{i=1}^{\hat{m}} w_i s_i(x), \tag{23}$$

where $w_i = \hat{g}(x_{i-1}) - \hat{g}(x_i)$ and $s_i(x) = H(x_i - x)$, with H being the Heaviside step function (i.e., 1 when its argument is non-negative, 0 otherwise). Notice that $\sum_i w_i = \hat{g}(0) - \hat{g}(1) = 1$. We will now construct our $g_{\hat{\mu}}$ as a mixture of measures $g_1, \ldots, g_{\hat{m}}$ each designed to approximate one of the step functions $H(x_i - x)$, and each a k-of-N measure. With $N = \lceil \frac{512\Delta^3}{\epsilon^3} \rceil$ and $K_i = \lceil x_i N \rceil$,

we choose $g_i(x) = 1 - I_x(K_i, N - K_i)$.

$$|g_i - s_i| = \int_0^1 |g_i(x) - s_i(x)| \tag{24}$$
$$= \int_0^1 |1 - I_x(K_i, N - K_i) - H(x_i - x)|\, dx \tag{25}$$
$$= \int_0^1 |H(x - x_i) - I_x(K_i, N - K_i)|\, dx \tag{26}$$

Let $\hat{x}_i = \frac{K_i}{N}$, then by the triangle inequality,

$$\leq \int_0^1 |H(x - \hat{x}_i) - I_x(K_i, N - K_i)| + |H(x - x_i) - H(x - \hat{x}_i)|\, dx \tag{27}$$
$$= \int_0^1 |H(x - \hat{x}_i) - I_x(K_i, N - K_i)|\, dx + \int_0^1 |H(x - x_i) - H(x - \hat{x}_i)|\, dx \tag{28}$$

The second integral can easily be bounded by $\frac{1}{N}$, and so we focus on the first. Let σ^2 be the variance of a Beta distributed random variable with parameters $(K_i, N - K_i)$ and let $c > 0$ be an arbitrary value.

$$\int_0^1 |H(x - \hat{x}_i) - I_x(K_i, N - K_i)|\, dx \tag{29}$$
$$= \int_0^{\hat{x}_i} |H(x - \hat{x}_i) - I_x(K_i, N - K_i)|\, dx + \int_{\hat{x}_i}^1 |H(x - \hat{x}_i) - I_x(K_i, N - K_i)|\, dx \tag{30}$$
$$= \int_0^{\hat{x}_i} I_x(K_i, N - K_i)\, dx + \int_{\hat{x}_i}^1 1 - I_x(K_i, N - K_i)\, dx \tag{31}$$
$$= \int_0^{\hat{x}_i - c\sigma} I_x(K_i, N - K_i)\, dx + \int_{\hat{x}_i - c\sigma}^{\hat{x}_i} I_x(K_i, N - K_i)\, dx$$
$$+ \int_{\hat{x}_i}^{\hat{x}_i + c\sigma} 1 - I_x(K_i, N - K_i)\, dx + \int_{\hat{x}_i + c\sigma}^1 1 - I_x(K_i, N - K_i)\, dx \tag{32}$$
$$\leq \int_0^{\hat{x}_i - c\sigma} I_x(K_i, N - K_i)\, dx + \int_{\hat{x}_i + c\sigma}^1 1 - I_x(K_i, N - K_i)\, dx$$
$$+ c\sigma I_{\hat{x}_i}(K_i, N - K_i) + c\sigma(1 - I_{\hat{x}_i}(K_i, N - K_i)) \tag{33}$$
$$= c\sigma + \int_0^{\hat{x}_i - c\sigma} I_x(K_i, N - K_i)\, dx + \int_{\hat{x}_i + c\sigma}^1 1 - I_x(K_i, N - K_i)\, dx \tag{34}$$

The remaining two integrals are simply the probability a Beta-distributed random variable is at least c standard deviations from its mean. By Chebyshev's inequality we have,

$$\int_0^{\hat{x}_i - c\sigma} I_x(K_i, N - K_i)\, dx + \int_{\hat{x}_i + c\sigma}^1 1 - I_x(K_i, N - K_i)\, dx \leq \frac{1}{c^2} \tag{35}$$

We can bound the variance of our Beta distribution by $\sigma^2 \leq \frac{1}{4N}$ and by choosing $c = N^{\frac{1}{6}}$ and putting all of the pieces together we have,

$$|g_i - s_i| \leq \frac{1}{N} + c\sigma + \frac{1}{c^2} \tag{36}$$

$$\leq \frac{1}{N} + \frac{N^{\frac{1}{6}}}{2N^{\frac{1}{2}}} + \frac{1}{N^{\frac{1}{3}}} \tag{37}$$

$$= \frac{1}{N} + \frac{1}{2N^{\frac{1}{3}}} + \frac{1}{N^{\frac{1}{3}}} \tag{38}$$

$$\leq \frac{1}{2N^{\frac{1}{3}}} + \frac{1}{2N^{\frac{1}{3}}} + \frac{1}{N^{\frac{1}{3}}} \tag{39}$$

$$= \frac{2}{N^{\frac{1}{3}}} \tag{40}$$

$$\leq 2\left(\frac{\epsilon^3}{512\Delta^3}\right)^{\frac{1}{3}} = \frac{\epsilon}{4\Delta} \tag{41}$$

So,

$$|g_{\hat{\mu}} - \hat{g}| = \int_0^1 |g_{\hat{\mu}}(x) - \hat{g}(x)|\, dx \tag{42}$$

$$= \int_0^1 \left| \left(\sum_i w_i g_i(x)\right) - \left(\sum_i w_i s_i(x)\right) \right| dx \tag{43}$$

$$= \int_0^1 \left| \sum_i w_i \left(g_i(x) - s_i(x)\right) \right| dx \tag{44}$$

$$\leq \int_0^1 \sum_i w_i \left|g_i(x) - s_i(x)\right|\, dx \tag{45}$$

$$\leq \sum_i w_i \int_0^1 \left|g_i(x) - s_i(x)\right|\, dx \tag{46}$$

$$\leq \sum_i w_i \frac{\epsilon}{4\Delta} = \frac{\epsilon}{4\Delta} \tag{47}$$

Thus, $|g_{\hat{\mu}} - g_{\mu}| \leq \frac{\epsilon}{2\Delta}$, which is what we set out to prove. Furthermore, $\hat{m}$ and N are polynomial in the quantities L, Δ, m, and $\frac{1}{\epsilon}$.

□

C The Diabetes Management Task

Our simplified diabetes management MDP is depicted in Figure 4. States in the diabetes management task are a combination of blood glucose level (low, medium, and high) and meal size (small, regular, and large). For simplicity, we do not simulate meal-size separately from blood glucose level even though they are highly tied to one another, i.e. patients who feel faint general sense that they have low blood sugar and choose to eat a bigger meal. Instead, our problem set up assumes that patients do not get to choose their meals and that insulin actions are the only thing determining transitions between states, i.e. that injecting some level of insulin affects not only how much blood glucose a patient has but also the meal they are to eat in the next step. A future task is to improve the diabetes management task to more accurately reflect reality.

Rewards depend only on blood glucose state and do not depend on insulin actions. They are drawn from Normal distributions: Low $\sim \mathcal{N}(5, 4)$, Med $\sim \mathcal{N}(6, 3)$, and High $\sim \mathcal{N}(5, 2)$. The values chosen loosely reflect what is desired by experts in the medical community. Medium blood glucose is ideal and yields the highest reward in expectation. A high-glucose state provides slightly lower rewards because a patient with higher blood glucose is in stable condition in the short-term, but may experience detrimental effects in the long-term. A low-glucose state is potentially very detrimental

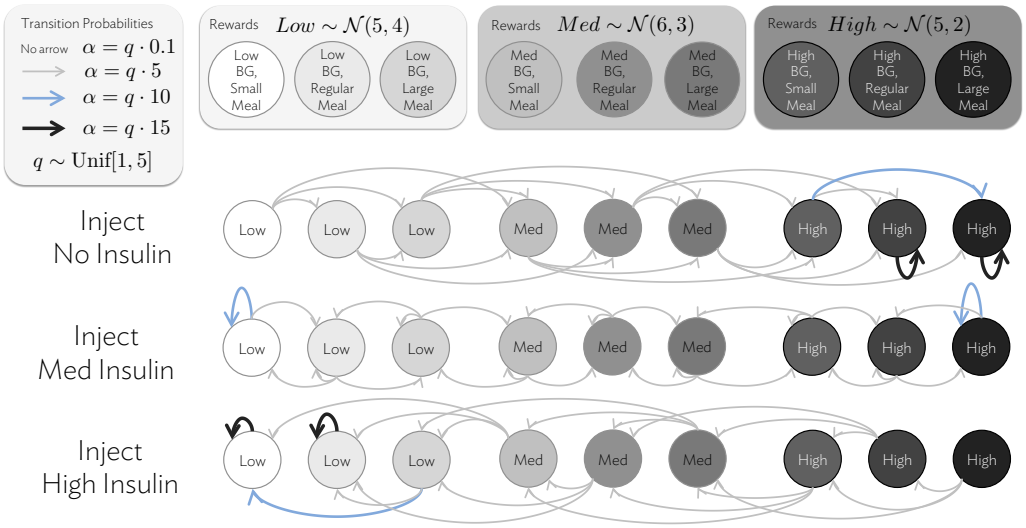

Figure 4: Transition probabilities between the states are shown as arrows of increasing width and darkness. The states (circles) get darker as we go toward high blood glucose states.

to the patient, as reflected in the high variance of the low-glucose reward distribution. We note that any choice of distribution could be used- Weibull, lognormal, uniform, and that Gaussian's were chosen because they were understandable and yielded interesting results.

Transitions are drawn from Dirichlet distributions. Figure 4 depicts transitions for the three levels of insulin. In general, patients trend toward higher blood glucose states after injecting no insulin. Patients generally stay within their current blood glucose level after injecting a moderate level of insulin, and they trend toward lower blood glucose states after injecting a high level of insulin. We scaled the Dirichlet parameters, α, with a multiplicative factor drawn from a uniform distribution $q \sim \text{Unif}[1, 5]$. When q is small, the transitions are more likely to take patients to unexpected states, a scenario that more adequately models child diabetics who often have unpredictable blood glucose fluctuations. When q is large, transitions more closely follow the depicted trends, a phenomenon more often observed with more stable adult diabetics.

D Computing k-of-N policies with CFR-BR

Algorithm 1 computes the k-of-N optimal policy, π^*, given N, k, and the number of CFR iterations T^*. The `EvaluatePolicyOnMDP` (line 6), `RegretUpdate` (line 11) and `PolicyUpdate` (line 12) steps differ when computing optimal policies under an imperfect recall assumption versus a perfect recall assumption as described below.

D.1 Imperfect recall

Under an assumption of reward uncertainty, the sequence of past states and actions is not informative and we find Markovian policies, which can be represented in tables size $O(|S|H)$. For the diabetes problem, there are $|S| = 9$ states, and $H = 3$ is the time horizon indexed by h. When $h = 0$, the agent is in end-game, and when $h = 3$, the agent is in the initial state as sampled from the initial state distribution $P(s_{\text{init}})$

Policy Evaluation. Given a policy π and an MDP with fixed reward and transition parameters R and P, we compute the state-action values $Q(s, a, h)$ and $v^{\pi}_{\mathcal{M}}$.

Initialize the state-action value function at end-game (terminal) states to be zero, i.e. $Q(s, a, h = 0) = 0$. For all non-terminal states, i.e. $\forall h \in (0, H]$, we use the following dynamic programming equation to compute the state-action value function:

Algorithm 1: CFR-BR for optimizing k-of-N percentile measures objectives

Data: N,k,T^*
Result: π^*
regret$^{t=0} = 0$;
$\pi^{t=0} = \frac{1}{|A|}$;
for $t = 1$ **to** T^* **do**
  for $i = 1$ **to** N **do**
    $\mathcal{M}_i \leftarrow$ `GenerateDiabetesMDP`;
    $v^{\pi}_{\mathcal{M}_i} \leftarrow$ `EvaluatePolicyOnMDP`;
  `SortMDPsOn`($v^{\pi}_{\mathcal{M}}$);
  $\mathcal{M}_{k\text{-of-}N} =$ `ChanceSampleFromBottom`(k);
  $v_{k\text{-of-}N} = v^{\pi}_{\mathcal{M}_{k\text{-of-}N}}$;
  $Q_{k\text{-of-}N} = \arg v_{k\text{-of-}N}$;
  regret$^t =$ `RegretUpdate`($Q_{k\text{-of-}N}$);
  $\pi^t =$ `PolicyUpdate`(regret^t);
$\pi^* = \pi^{T^*}$

$$Q(s,a,h) = R(s) + \sum_{s'}\sum_{a'} P_a(s,s')\pi(s',a',h-1)Q(s',a',h-1) \tag{48}$$

The value of employing policy π in and MDP $\mathcal{M}$ is given by

$$v^{\pi}_{\mathcal{M}} = \sum_{s}\sum_{a} P(s_{\text{init}})\pi(s,a,H)Q(s,a,H) \tag{49}$$

Regret Update. After sorting $v^{\pi}_{\mathcal{M}}$ for all N MDPs, chance samples one MDP $\mathcal{M}_{k\text{-of-}N}$ from the bottom k. Regrets are updated according to the k-of-N state-action value function $Q_{k\text{-of-}N}$

$$\text{regret}^t(s,a,h) = \text{regret}^{t-1}(s,a,h) + \left(Q_{k\text{-of-}N}(s,a,h) - \sum_{a'\in A}\pi^t(s,a',h)Q_{k\text{-of-}N}(s,a',h)\right) \tag{50}$$

Policy Update.

$$\pi(s,a,h)^{t+1} = \frac{(\text{regret}^t(s,a,h))^+}{\sum_a(\text{regret}^t(s,a,h))^+} \tag{51}$$

where $f^+ = \max(f,0)$.

D.2 Perfect recall

Under an assumption of transition uncertainty, policies depend on entire histories of past states and actions. As a result, the number of information sets (i.e., decision points) in an optimal policy is $|\mathcal{I}_1| = |S|((|S||A|)^H - 1)/(|S||A| - 1)$, and so polynomial in the number of states and actions for any fixed horizon, but exponential in the horizon itself. The diabetes problem has $|X| = |S|((|S||A|)^H - 1)/(|S||A| - 1) = 6813$ non-terminal histories, denoted as x, and $|Z| = |S|(|S||A|^H) = 177147$ terminal leaf nodes, denoted as z.

Policy Evaluation. Given a policy π and an MDP with fixed reward and transition parameters R and P as drawn from their respective distributions, we compute the state-action values $Q(x,a)$ and $v^{\pi}_{\mathcal{M}}$.

We initialize the state-action value function at end-game (terminal) states (the leaf nodes in the game tree) z to be the terminal rewards, i.e. $Q(z,a) = R(z)$. For all non-terminal states, x, we compute:

$$Q(x,a) = \sum_{x'}\sum_{a'} P_a(x,x')\pi(x',a')Q(x',a') \tag{52}$$

where x is a parent node of x' in the game tree, i.e. if x' is the sequence $(s_0, a_5, s_1, a_4, s_2)$, then the parent x is the sequence (s_0, a_5, s_1).

$$v^{\pi}_{\mathcal{M}} = \sum_{x_{\text{init}}} \sum_{a} P(x_{\text{init}}) \pi(x_{\text{init}}, a) Q(x_{\text{init}}, a) \tag{53}$$

The regret and policy updates are directly analogous to the one under an imperfect recall assumption.

Regret Update.

$$\text{regret}^t(x, a) = \text{regret}^{t-1}(x, a) + \left(Q_{k\text{-of-}N}(x, a) - \sum_{a'} \pi^t(x, a') Q_{k\text{-of-}N}(x, a') \right) \tag{54}$$

Policy Update.

$$\pi^{t+1}(x, a) = \frac{(\text{regret}^t(x, a))^+}{\sum_a (\text{regret}^t(x, a))^+} \tag{55}$$