Count-Based Frequency Estimation with Bounded Memory

Supplemental

Marc G. Bellemare

May 1, 2015

1 Notation

Let μ be a probability distribution over an alphabet $\mathcal{X}$, with Shannon entropy $H(\mu)$. For a set $S \subseteq \mathcal{X}$ we write $\mu(S) := \sum_{x \in S} \mu(x)$. For a random variable X taking values in the discrete alphabet $\mathcal{X}$ and F an event in the σ-algebra of X, we write $\mathbf{E}_\mu[X; F] := \mathbf{E}_\mu[X \mathbb{I}_{[F]}]$. Whenever unambiguous we write $\mathrm{KL}(\cdot \,\|\, \cdot) := \mathrm{KL}_1(\cdot \,\|\, \cdot)$. Finally, for an integer $i \in \mathbb{N}$ we write $[i] := \{1, \ldots, i\}$. We refer the reader to the main text for the full list of terms.

Throughout, we make use of the following information-theoretic property: for any discrete random variable X over $\mathcal{X}$ we have (Cover and Thomas, 1991),

$$\mathrm{KL}(X \,\|\, \mathscr{U}_\mathcal{X}) = \log |\mathcal{X}| - H(X). \tag{1}$$

Lemma 1. *Let $\mathcal{M}$ be the class of budget multinomial distributions. Let μ be a memoryless stationary source and $K \in \mathbb{N}$. Then the model $\rho^*_\mu \in \mathcal{M}$ which minimizes $KL(\mu \,\|\, \cdot)$ is*

$$\mathcal{Z}^* = \operatorname*{arg\,min}_{\mathcal{Z}: |\mathcal{Z}| = K} \left[\mu(\mathcal{X} \setminus \mathcal{Z}) KL(\mu_{\mathcal{X} \setminus \mathcal{Z}} \,\|\, \mathscr{U}_{\mathcal{X} \setminus \mathcal{Z}}) \right]$$
$$\theta^*(x) = \mu(x) \quad \forall x \in \mathcal{Z}^* \qquad \theta^*_0 = \mu(\mathcal{X} \setminus \mathcal{Z}^*).$$

Proof. First, we fix $\mathcal{Z}$ and derive the optimal parameters for this $\mathcal{Z}$. Second, we use these optimal parameters to obtain $\mathcal{Z}^*$. We begin by expanding the KL divergence of $\rho(x) := \rho(x\,;\, \mathcal{Z}, \theta, \theta_0)$ from μ:

$$\begin{aligned} J(\mathcal{Z}, \theta, \theta_0) := \mathrm{KL}(\mu \,\|\, \rho) &= \mathop{\mathbf{E}}_{x \sim \mu} \left[\log \frac{\mu(x)}{\rho(x)} \right] \\ &= \mathop{\mathbf{E}}_{x \sim \mu} \left[\log \frac{\mu(x)}{\rho(x)} \,;\, x \in \mathcal{Z} \right] + \mathop{\mathbf{E}}_{x \sim \mu} \left[\log \frac{\mu(x)}{\rho(x)} \,;\, x \notin \mathcal{Z} \right] \end{aligned}$$

We now define the Lagrangian

$$J(\mathcal{Z}, \theta, \theta_0, \lambda) := J(\mathcal{Z}, \theta, \theta_0) + \lambda \left(\sum_{x \in \mathcal{X}} \rho(x) - 1 \right).$$

Let $\theta := [\theta_1, \ldots, \theta_K]$, and fix $\mathcal{Z} := \{x^1, \ldots, x^K\}$. For $i \in [K]$, we have

$$\begin{aligned}\frac{\partial}{\partial \theta_i} J(\mathcal{Z}, \theta, \theta_0, \lambda) &= \frac{\partial}{\partial \theta_i} \mathop{\mathbf{E}}_{x\sim\mu}\left[\log \frac{\mu(x)}{\rho(x)}\,;\, x \in \mathcal{Z}\right] + \lambda \\ &= -\frac{\mu(x^i)}{\rho(x^i)} + \lambda.\end{aligned}$$

Recall that $\rho(x) = \frac{\theta_0}{|\mathcal{X}\setminus\mathcal{Z}|}$ for $x \notin \mathcal{Z}$. We thus have

$$\begin{aligned}\frac{\partial}{\partial \theta_0} J(\mathcal{Z}, \theta, \theta_0, \lambda) &= \frac{\partial}{\partial \theta_0} \mathop{\mathbf{E}}_{x\sim\mu}\left[\log \frac{\mu(x)}{\rho(x)}\,;\, x \notin \mathcal{Z}\right] + \lambda \sum_{x\in\mathcal{X}} \mathbb{I}_{[x\notin\mathcal{Z}]} \\ &= -\sum_{x\notin\mathcal{Z}} \mu(x)\frac{1}{\theta_0} + \lambda|\mathcal{X}\setminus\mathcal{Z}| \\ &= -\frac{\mu(\mathcal{X}\setminus\mathcal{Z})}{\theta_0} + \lambda|\mathcal{X}\setminus\mathcal{Z}|\end{aligned}$$

Setting $\frac{\partial}{\partial\theta_i} J(\mathcal{Z}, \theta, \theta_0, \lambda) = 0$ for $i \in [K] \cup \{0\}$, we obtain

$$\theta_i = \lambda^{-1}\mu(x^i) \qquad \theta_0 = \lambda^{-1}\frac{\mu(\mathcal{X}\setminus\mathcal{Z})}{|\mathcal{X}\setminus\mathcal{Z}|}$$

Finally, from $\sum_{x\in\mathcal{X}} \rho(x) = 1$ we find that $\lambda = 1$. For a fixed $\mathcal{Z}$, the unique optimal parameters are thus

$$\theta_i^{\mathcal{Z}} = \mu(x^i) \qquad \theta_0^{\mathcal{Z}} = \frac{\mu(\mathcal{X}\setminus\mathcal{Z})}{|\mathcal{X}\setminus\mathcal{Z}|} \tag{2}$$

We plug these values into the definition of $J(\mathcal{Z}, \theta, \theta_0)$:

$$\begin{aligned}J(\mathcal{Z}, \theta, \theta_0) &= \mathop{\mathbf{E}}_{x\sim\mu}\left[\log \frac{\mu(x)}{\rho(x)}\,;\, x \in \mathcal{Z}\right] + \mathop{\mathbf{E}}_{x\sim\mu}\left[\log \frac{\mu(x)}{\rho(x)}\,;\, x \notin \mathcal{Z}\right] \\ &= \mathop{\mathbf{E}}_{x\sim\mu}\left[\log \frac{\mu(x)}{\mu(x)}\,;\, x \in \mathcal{Z}\right] + \mathop{\mathbf{E}}_{x\sim\mu}\left[\log \frac{\mu(x)}{\theta_0^{\mathcal{Z}}}\,;\, x \notin \mathcal{Z}\right] \\ &= \mathop{\mathbf{E}}_{x\sim\mu}\left[\log \frac{\mu(x)\mu(\mathcal{X}\setminus\mathcal{Z})^{-1}}{|\mathcal{X}\setminus\mathcal{Z}|^{-1}}\,;\, x \notin \mathcal{Z}\right] \\ &= \mathop{\mathbf{E}}_{x\sim\mu}\left[\log \frac{\mu_{\mathcal{X}\setminus\mathcal{Z}}(x)}{\mathscr{U}_{\mathcal{X}\setminus\mathcal{Z}}(x)}\,;\, x \notin \mathcal{Z}\right] \\ &= \mu(\mathcal{X}\setminus\mathcal{Z})\mathrm{KL}(\mu_{\mathcal{X}\setminus\mathcal{Z}} \,\|\, \mathscr{U}_{\mathcal{X}\setminus\mathcal{Z}}).\end{aligned}$$

The model in $\mathcal{M}$ which minimizes $\mathrm{KL}(\mu\,\|\,\cdot)$ is thus

$$\mathcal{Z}^* = \mathop{\arg\min}_{\mathcal{Z}:|\mathcal{Z}|=K} \left[\mu(\mathcal{X}\setminus\mathcal{Z})\mathrm{KL}(\mu_{\mathcal{X}\setminus\mathcal{Z}} \,\|\, \mathscr{U}_{\mathcal{X}\setminus\mathcal{Z}})\right],$$

with parameters given by Equation 2. □

Lemma 2. *Let $\mathcal{Z}^*$ be the subalphabet associated with the model $\rho_\mu^* \in \mathcal{M}$ minimizing $KL(\mu\,\|\,\cdot)$. Then $\mathcal{Z}^* = H \cup L$, where $H, L \subseteq \mathcal{X}$ are such that*

- *for all* $x \in H, y \in \mathcal{X} \setminus H$, $\mu(x) \geq \mu(y)$, *and*
- *for all* $x \in L, y \in \mathcal{X} \setminus L$, $\mu(x) \leq \mu(y)$.

Proof. From Lemma 1, we know the optimal model has subalphabet

$$\mathcal{Z}^* = \operatorname*{arg\,min}_{\mathcal{Z}:|\mathcal{Z}|=K} \left[\mu(\mathcal{X} \setminus \mathcal{Z})\mathrm{KL}(\mu_{\mathcal{X}\setminus\mathcal{Z}} \,\|\, \mathscr{U}_{\mathcal{X}\setminus\mathcal{Z}})\right].$$

Let $\mathcal{X} := \{x^1, x^2, \ldots, x^m\}$ be such that $\mu(x^1) \geq \mu(x^2) \geq \cdots \geq \mu(x^i) \geq \ldots \mu(x^m)$. For simplicity we consider the case when $K = 1$, and show that

$$\mu(\mathcal{X} \setminus \mathcal{Z})\mathrm{KL}(\mu_{\mathcal{X}\setminus\mathcal{Z}} \,\|\, \mathscr{U}_{\mathcal{X}\setminus\mathcal{Z}})$$

is concave over $i = 1, \ldots, m$. For $K = 1$ and a fixed choice of x^i, the above can be rewritten as

$$\begin{aligned}
\mu(\mathcal{X} \setminus \mathcal{Z})\mathrm{KL}(\mu_{\mathcal{X}\setminus\mathcal{Z}} \,\|\, \mathscr{U}_{\mathcal{X}\setminus\mathcal{Z}}) &= (1 - \mu(x^i))\mathrm{KL}(\mu_{\mathcal{X}\setminus\mathcal{Z}} \,\|\, \mathscr{U}_{\mathcal{X}\setminus\mathcal{Z}}) \\
&= (1 - \mu(x^i))\left[\log|\mathcal{X} \setminus \mathcal{Z}| - H(\mu_{\mathcal{X}\setminus\mathcal{Z}})\right] \\
&= (1 - \mu(x^i))\left[\log|\mathcal{X} \setminus \mathcal{Z}| + \sum_{x \neq x^i} \frac{\mu(x)}{\mu(\mathcal{X} \setminus \mathcal{Z})} \log\left(\frac{\mu(x)}{\mu(\mathcal{X} \setminus \mathcal{Z})}\right)\right] \\
&= (1 - \mu(x^i))\log|\mathcal{X} \setminus \mathcal{Z}| + \sum_{x \neq x^i} \mu(x)\log\mu(x) - \sum_{x \neq x^i} \log\mu(\mathcal{X} \setminus \mathcal{Z}) \\
&= (1 - \mu(x^i))\log|\mathcal{X} \setminus \mathcal{Z}| - H(\mu) \\
&\qquad - \mu(x^i)\log\mu(x^i) - (1 - \mu(x^i))\log(1 - \mu(x^i)) \\
&= (1 - \mu(x^i))\log|\mathcal{X} \setminus \mathcal{Z}| - H(\mu) + H(\mu(x^i)),
\end{aligned}$$

where with some abuse of notation we use $H(\mu(x^i))$ to denote the entropy of a Bernoulli distribution with parameter $\mu(x^i)$. This entropy is concave in $\mu(x^i)$, and summary examination shows that the whole function is also concave in $\mu(x^i)$. It must be therefore be that the minimum is achieved by either $\mu(x^1)$ or $\mu(x^m)$, i.e. by removing either the least or most frequent element from $\mathcal{X}$. The case $K > 1$ follows by similarly relating our objective function to the entropy of a multinomial distribution with parameters $\mu(x^{i_1}), \ldots, \mu(x^{i_m}), 1 - \mu(\mathcal{X} \setminus \mathcal{Z})$. □

Lemma 3. *Let* $\mathcal{M}$ *be the class of budget multinomial distributions. Let* μ *be a memoryless stationary source and* $K \in \mathbb{N}$. *Let* $\rho^* := \rho^*_\mu \in \mathcal{M}$ *be the model with parameter* $\mathcal{Z}^*$ *minimizing* $KL(\mu \,\|\, \cdot)$, *and let* $\hat{\rho} := \hat{\rho}_\mu \in \mathcal{M}$ *be the model defined according to*

$$\hat{\mathcal{Z}} := \operatorname*{arg\,max}_{\mathcal{Z}:|\mathcal{Z}|=K} \mu(\mathcal{Z})$$

with parameters as per Lemma 1. Then

$$KL(\mu \,\|\, \hat{\rho}) - KL(\mu \,\|\, \rho^*) \leq \frac{1}{e} + \mu(\mathcal{Z}^* \setminus \hat{\mathcal{Z}})\log K \leq \frac{1}{e} + \frac{K}{|\mathcal{X}|}\log K.$$

Furthermore, there exists a source μ *for which*

$$KL(\mu \,\|\, \hat{\rho}) - KL(\mu \,\|\, \rho^*) \geq \frac{1}{e}$$

Proof. Let $|\mathcal{X}| = m$ and as before let $\mathcal{X} := \{x^1, \ldots, x^m\}$ such that $\mu(x^1) \geq \mu(x^2) \geq \mu(x^m)$. Without loss of generality, we consider the case where $\mathcal{Z}^* := \{x^{m-K+1}, \ldots, x^m\}$. We partition our alphabet into three sets, $A := \hat{\mathcal{Z}}$, $C := \mathcal{Z}^*$, and $B := \mathcal{X} \setminus \{A \cup B\}$ (Figure 1) and let $L := |B| + K = |\mathcal{X}| - K$. We begin with a few properties which follow from our arrangement of $\mathcal{X}$ in order of decreasing

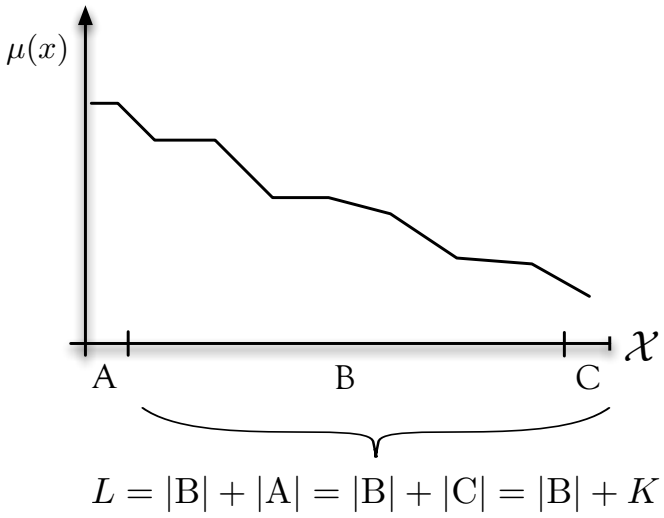

Figure 1: Probability distribution over $\mathcal{X}$ with symbols arranged by decreasing probability. Here $A := \hat{\mathcal{Z}}$, $C := \mathcal{Z}^*$ and $B := \mathcal{X} \setminus \{A \cup C\}$.

probability:

$$\rho^*(x) = \frac{\mu(A) + \mu(B)}{L} \; \forall x \in A \cup B \qquad \hat{\rho}(x) = \frac{\mu(B) + \mu(C)}{L} \; \forall x \in B \cup C$$

$$\hat{\rho}(x) \geq \rho^*(x) \; \forall x \in A \qquad \frac{\mu(B) + \mu(C)}{L} \geq \frac{\mu(C)}{K}$$

Observe that both models assign uniform probabilities over B, but may differ significantly over A and C. We consider the difference in their KL divergence from μ:

$$\mathrm{KL}(\mu \,\|\, \hat{\rho}) - \mathrm{KL}(\mu \,\|\, \rho^*) = \sum_{S \in \{A,B,C\}} \mathop{\mathbf{E}}_{x \sim \mu} \left[\log \frac{\rho^*(x)}{\hat{\rho}(x)} \,;\, x \in S \right] \tag{3}$$

Since $\hat{\rho}(x) \geq \rho^*(x)$ for all $x \in \mathcal{A}$, we know only the expectations for sets B and C may be positive in Equation 3. We first consider the B term:

$$\begin{aligned}
\mathop{\mathbf{E}}_{x \sim \mu} \left[\log \frac{\rho^*(x)}{\hat{\rho}(x)} \,;\, x \in B \right] &= \mathop{\mathbf{E}}_{x \sim \mu} \left[\log \frac{\mu(A) + \mu(B)}{\mu(B) + \mu(C)} \,;\, x \in B \right] \\
&\leq \mathop{\mathbf{E}}_{x \sim \mu} \left[\log \frac{1}{\mu(B)} \,;\, x \in B \right] \\
&\leq -\mu(B) \log(\mu(B)) \\
&\leq \frac{1}{e},
\end{aligned}$$

where the last inequality follows from finding the maximum of $-x \log x$ on $[0, 1]$. Turning our attention to the C term, we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{x\sim\mu}\left[\log \frac{\rho^*(x)}{\hat{\rho}(x)}\,;\, x \in C\right] &= \mathop{\mathbf{E}}_{x\sim\mu}\left[\log \frac{\mu(x)}{(\mu(B)+\mu(C))/L}\,;\, x \in C\right] \\
&\leq \mathop{\mathbf{E}}_{x\sim\mu}\left[\log \frac{\mu(x)}{\mu(C)/K}\,;\, x \in C\right] \\
&\leq \mu(C) \mathop{\mathbf{E}}_{x\sim\mu_C}\left[\log \frac{\mu_C(x)}{1/K}\right] \\
&= \mu(C)\mathrm{KL}(\mu_C \,\|\, \mathscr{U}_C) \\
&\leq \mu(C) \log K,
\end{aligned}$$

where for the second inequality we used $\mu_C(x) = \mu(x)/\mu(C)$ on $x \in C$, and the last line follows from Equation 1. Since C contains the K symbols with smallest probabilities, we know by the pigeonhole principle that $\mu(C) \leq \frac{K}{|\mathcal{X}|}$; noting that in the general case $C = \mathcal{Z}^* \setminus \hat{\mathcal{Z}}$, the upper bound now follows by combining the terms.

To prove the lower bound, we simply construct a source μ for which $\mu(x) = c$ for all $x \in A \cup B$, and $\mu(x) = \epsilon < c$ for $x \in C$. In this case, $\mathrm{KL}(\mu \,\|\, \rho^*) = 0$ since ρ^* correctly assigns a uniform probability to all symbol in $A \cup B$; the $\frac{1}{e}$ lower bound follows by maximizing $-\mu(B) \log \mu(B)$ as before. □

Theorem 1. *When Algorithm 1 (K-distinct reservoir sampling) terminates, its summary $\mathcal{S}$ is the K-concise summary of a permutation $\tilde{x}_{1:n}$ sampled uniformly at random from $\mathcal{P}(x_{1:n})$.*

Proof. Let $(r_t : t \in \mathbb{N})$ be the sequence of random integers drawn by Algorithm 1, and $(w^t_{1:t} : t \in \mathbb{N})$ the sequence of permutations of $x_{1:n}$ induced by (r_t), with $w^t_{1:t} := w^t_1 \dots w^t_t \in \mathcal{X}^t$. Note that this sequence of permutations does not depend on the rest of Algorithm 1.

Our proof goes by induction. At time $t = 1$ we insert x_1 with $d_{t,1} = 1 = \tau_t(x_1)$; clearly $\mathcal{S}_1 = \langle (x_1, 1) \rangle$ is the 1-concise summary of the unique permutation $w^1_{1:1} = x_1$. Now assume that $\mathcal{S}_{t-1} := \langle (y_{t-1,i}, d_{t-1,i}) \rangle_{i=1}^{K'}$ is the K-concise summary of $w^{t-1}_{1:t-1}$, and without loss of generality assume $K' = K$. For clarity of exposition let $r := r_t$, $I := I_{t-1}(x_t)$, and $J := J_{t-1}(r)$; further, we write $D_{t,i} := \sum_{j=1}^{i-1} d_{t,j}$ (see main text for definitions).

Observe that r is the position of x_t in $w^t_{1:t}$ and in particular $w^t_{1:t} = w^{t-1}_{1:r} x_t w^{t-1}_{r+1:t-1}$. By assumption, for all $i \leq K$ we have $D_{t-1,i} = \tau(y_i, w^{t-1}_{1:t-1}) - 1$, and furthermore $\mathbf{y}_{t-1} := \mathbf{y}(\mathcal{S}_{t-1}) = \phi_K(w^{t-1}_{1:t-1})$. We need to show that $D_{t,i} = \tau(y_i, w^t_{1:t}) - 1$ for all $i \leq K$ and $\mathbf{y}_t = \phi_K(w^t_{1:t})$.

Consider first the case when x_t occurs within $w_{1:r}$, i.e. $I < J$. Then the first occurrence of x_t, $\tau(x_t, w^t_{1:t})$, is the same as $\tau(x_t, w^{t-1}_{1:t-1})$. More generally, for all $i < J$ we have $\tau(y_{t-1,i}, w^t_{1:t}) = \tau(y_{t-1,i}, w^{t-1}_{1:t-1})$, while for $i \geq J$ we have $\tau(y_{t-1,i}, w^t_{1:t}) = \tau(y_{t-1,i}, w^{t-1}_{1:t-1}) + 1$. Furthermore, $\phi_K(w^t_{1:t}) = \phi_K(w^{t-1}_{1:t-1})$. In this case we set $d_{t,J-1} \leftarrow d_{t-1,J-1} + 1$, such that for all $i \geq J$ we have $D_{t,i} = \sum_{j=1}^{i-1} d_{t,j} = D_{t-1,i} = \tau(y_{t-1,i}, w^t_{1:t}) - 1$, as desired.

If x_t does not occur within $w_{1:r}$, we must insert it within our K-concise summary, provided that $x_t \in \phi_K(w_{1:r} x_t)$, i.e. fewer than K distinct symbols occur in $w_{1:r}$. Again, for all $i < J$, we have $\tau(y_{t-1,i}, w^t_{1:t}) = \tau(y_{t-1,i}, w^{t-1}_{1:t-1})$, while for $i > J$ we have $\tau(y_{t-1,i}, w^t_{1:t}) = \tau(y_{t-1,i}, w^{t-1}_{1:t-1}) + 1$. However, the time of first occurrence of x_t now naturally becomes r, and $\phi_K(w^t_{1:t}) \neq \phi_K(w^{t-1}_{1:t-1})$. In this case $\mathcal{S}_t$ is updated from $\mathcal{S}_{t-1}$ as follows. If x_t occurs in $\phi_K(w^{t-1}_{1:t-1})$ (at position I), we first

remove it from the reservoir and set $d_{t,I-1} = d_{t-1,I-1} + d_{t-1,I}$. We then INSERT x_t at position J with count $D_{t-1,J} - r + 1$, as per Algorithm 1; in particular, all elements at positions $i \geq J$ are moved forward one position, i.e. $y_{t,i+1} = y_{t-1,i}$. Following the insertion, the $D_{t,\cdot}$ terms become

$$\begin{aligned} D_{t,J} &= \sum_{j=1}^{J-1} d_{t,j} = \sum_{j=1}^{J-2} d_{t-1,j} + r - D_{t-1,J-1} = r = \tau(x_t, w_{1:t}^t) - 1, \\ D_{t,i} &= \sum_{j=1}^{i-1} d_{t,j} = D_{t,J} + D_{t-1,J} - r + 1 + \sum_{j=J}^{i-1} d_{t-1,j} = D_{t-1,i} + 1 = \tau(y_{t,i}, w_{1:t}^t) - 1 \; \forall i > J, \end{aligned}$$

as desired. Finally, we remove the $K^{th} + 1$ symbol. For $y_1 \dots y_K := \mathbf{y}_{t-1} = \phi_K(w_{1:t-1}^{t-1})$, then

$$\phi_K(w_{1:t}^t) = y_1 \dots y_{J-1} x_t y_J \dots y_{K-1},$$

which matches the contents of $\mathcal{S}_t$. Thus $\mathcal{S}_t$ contains the K-concise summary for the permutation $w_{1:t}^t$, and the result follows by induction. □

Lemma 4. *Let $(\mu_t : t \in \mathbb{N})$ be a sequence of probability distributions over $\mathcal{X}$ together forming a memoryless coding distribution $\mu(x_{1:n}) = \prod_{t=1}^n \mu_t(x_t)$. Let $x_{1:n}$ be a random string drawn from μ, and $\tilde{x}_{1:n}$ a permutation of $x_{1:n}$ drawn uniformly at random from $\mathcal{P}(x_{1:n})$. For any $x \in \mathcal{X}$, we have*

$$\Pr\{x \notin \phi_K(\tilde{x}_{1:n})\} \leq (1 - \min_{j \leq n} \mu_j(x))^K$$

Proof. Consider the permutation π on $[n]$ mapping $x_{1:n}$ to $\tilde{x}_{1:n}$ such that $\tilde{x}_i = x_{\pi(i)}$. If the random elements of $\phi_K(\tilde{x}_{1:n})$ are denoted $Y_1 \dots Y_K$, then

$$\begin{aligned} \Pr\{x \notin \phi_K(\tilde{x}_{1:n})\} &= \sum_{y_1, \dots, y_K \neq x} \Pr\{Y_1 = y_1, \dots, Y_K = y_K\} \\ &= \sum_{y_1, \dots, y_{K-1} \neq x} \Pr\{Y_1, \dots, Y_{K-1}\} \sum_{y_K \neq x} \Pr\{Y_K = y_k \,|\, Y_1, \dots, Y_{K-1}\}, \end{aligned}$$

where in the second line $Y_i = y_i$ is implied for conciseness. Now,

$$\sum_{y_K \neq x} \Pr\{Y_K = y_K \,|\, Y_1, \dots, Y_{K-1}\} = 1 - \Pr\{Y_K = x \,|\, Y_1, \dots Y_{K-1}\}$$

In general, the distribution over Y_K depends on the random time τ_K at which the K^{th} distinct symbol first occurs in $\tilde{x}_{1:n}$. For example, the first element $\tilde{x}_1$ is sampled from $\mu_{\pi(1)}$, the second from $\mu_{\pi(\tau_2)}$, etc. However, given $\tau_1 \dots \tau_{K-1}$, we can lower bound $\Pr\{Y_K = x \,|\, Y_1, \dots, Y_{K-1}\}$ as

$$\Pr\{Y_K = x \,|\, Y_1, \dots Y_{K-1}\} \;\geq\; \frac{\min_{j \leq n} \mu_j(x)}{1 - \sum_{j=1}^i \mu_{\pi(\tau_j)}(y_j)} \leq \min_{j \leq n} \mu_j(x).$$

Since the resulting probability is independent of $\tau_1 \dots \tau_{K-1}$, we can write

$$\begin{aligned} \Pr\{x \notin \phi_K(\tilde{x}_{1:n})\} &= \sum_{y_1, \dots, y_{K-1} \neq x} \Pr\{Y_1, \dots, Y_{K-1}\} \left(1 - \Pr\{Y_K = x \,|\, Y_1, \dots, Y_{K-1}\}\right) \\ &\leq (1 - \min_{j \leq n} \mu_j(x)) \sum_{y_1, \dots, y_{K-1} \neq x} \Pr\{Y_1, \dots, Y_{K-1}\} \end{aligned}$$

and by induction on K

$$\Pr\{x \notin \phi_K(\tilde{x}_{1:n})\} \leq (1 - \min_{j \leq n} \mu_j(x))^K.$$

Since this bound does not depend on π, the desired result follows. □

Lemma 5 (Lemma 3 in the main text)**.** *Let μ be a memoryless stationary source, let $\mathcal{B} \subseteq \mathcal{X}, \delta \in (0,1)$, and let $\upsilon(\mathcal{B}, \mu) := \min_{x \in \mathcal{B}} \mu(x)$. Let $\mathscr{Y} := (\mathcal{Y}_t : t \in \mathbb{N})$ be the sequence of alphabets $\mathcal{Y}_t := \mathcal{Y}(\mathcal{S}_t) \subseteq \mathcal{X}$ induced by running Algorithm 1 on a string $x_{1:n} \sim \mu$. Further let $G_n(x)$, $G_n(\mathcal{B})$ be the events defined as*

$$G_n(x) := \{\forall t \geq \tau(x, x_{1:n}), x \in \mathcal{Y}_t\} \qquad G_n(\mathcal{B}) := \bigcap_{x \in \mathcal{B}} G_n(x).$$

Then for $K \geq \upsilon(\mathcal{B}, \mu)^{-1} \log(|\mathcal{B}| n \delta^{-1})$, we have

$$\Pr\{G_n(\mathcal{B})\} \geq 1 - \delta.$$

Corollary 1. *Under the same conditions as Lemma 5, with probability $1 - \delta$ the Budget SAD contains the correct counts for all $x \in \mathcal{B}$, i.e. $c_{n, I_n(x)} = N_n(x)$.*

Proof (of Lemma 5). Let $\mathcal{A}_t := \{x \in \mathcal{X} : N_t(x) > 0\}$. By definition, Lemma 5 is trivially true whenever $\tau_n(x) = n + 1$. For x such that $\tau_n(x) \leq n$, we begin with

$$G_n(x) = \bigcap_{t=\tau_n(x)}^{n} (x \in \mathcal{Y}_t) = \bigcap_{t=1}^{n} (x \notin \mathcal{A}_t \cup x \in \mathcal{Y}_t).$$

Using De Morgan's law, we rewrite the above as

$$\neg G_n(x) = \bigcup_{t=1}^{n} (x \in \mathcal{A}_t \cap x \notin \mathcal{Y}_t) \subseteq \bigcup_{t=1}^{n} (x \notin \mathcal{Y}_t)$$

Recall that $\mathcal{Y}_t := \mathcal{Y}(\mathcal{S}_t)$ is the alphabet corresponding to our K-distinct sample of $x_{1:t}$. By Lemma 4 and the definition of $\upsilon(\mathcal{B}, \mu)$ we have, for any $x \in \mathcal{B}$,

$$\begin{aligned} \Pr\{x \notin \mathcal{Y}_t\} &\leq (1 - \mu(x))^K \\ &\leq (1 - \upsilon(\mathcal{B}, \mu))^K \\ &\leq e^{-\upsilon(\mathcal{B}, \mu) K} \\ &\leq e^{-\log \frac{|\mathcal{B}| n}{\delta}} = (|\mathcal{B}| n)^{-1} \delta, \end{aligned}$$

where we used the fact that $0 \leq \upsilon(\mathcal{B}, \mu) \leq 1$ (third line). From a union bound over $\mathcal{B}$ and $t = 1, \ldots, n$, we obtain

$$\Pr\{\neg G_n(\mathcal{B})\} \leq \sum_{x \in \mathcal{B}} \sum_{t=1}^{n} \Pr\{x \notin \mathcal{Y}_t\} \leq \delta$$

It follows that $G_n(\mathcal{B})$ must occur with probability at least $1 - \delta$. □

We now develop three lemmas needed to prove an expected redundancy bound for the Budget SAD. Two random sequences arise in our analysis of this expected redundancy: the sequence of symbols $x_{1:n} \sim \mu$ and the sequence of random insertions $r_{1:n}$, sampled according to $r_t \sim \mathscr{U}(\{0, \ldots, t-1\})$. These two sequences together induce a random sequence of K-distinct summaries, $(\mathcal{S}_t : t \in \mathbb{N})$, with $\mathcal{S}_t = [(y_{t,i}, d_{t,i}, c_{t,i}) \in \mathcal{X} \times \mathbb{N}^+ \times \mathbb{N}^+ : i \in [K]]$, which itself induces a sequence $(D_{t,i} : t \in \mathbb{N}, i \in [K])$ with $D_{t,i} := \sum_{j=1}^{i-1} d_{t,j}$.

Our redundancy bound depends on an augmented version of $\mathcal{Z}^*$ which contains $K > |\mathcal{Z}^*|$ elements. This augmented set $\mathcal{Z}_{\text{AUG}}$ is defined as

$$\mathcal{Z}_{\text{AUG}} := \arg\max_{\mathcal{B} \subseteq \mathcal{X}} \mu(\mathcal{B}).$$

Since $K > |\mathcal{Z}^*|$ and $\mathcal{Z}^*$ is composed of the most frequent symbols in μ, it follows that $\mathcal{Z}^* \subseteq \mathcal{Z}_{\text{AUG}}$. We begin by bounding the expected "length" of our K-distinct summary, $D_{t,K+1}$, which we then use to bound $\mathbf{E}[z_t]$, the expected number of discards. This latter quantity plays a critical role in the expected redundancy of the Budget SAD.

Lemma 6. *For any $t \in \mathbb{N}$, we have*

$$\mathop{\mathbf{E}}_{x_{1:n}, r_{1:n}} [D_{t,K+1}] \leq \min\left\{t - 1, K(1 - \mu(\mathcal{Z}_{\text{AUG}}))^{-1} - 1\right\}$$

Proof. As previously, let $\tilde{x}_{1:n} \sim \mathscr{U}(\mathcal{P}(x_{1:n}))$ denote the random permutation of $x_{1:n}$ induced by $r_{1:n}$, and let $\mathbf{y} := y_1 \ldots y_K := \phi_K(\tilde{x}_{1:n})$. Further let $\mathscr{G}(p)$ be the geometric distribution with mean p^{-1}. Recall that $D_{t,K+1}$ is the first occurrence of the first $x \notin \mathbf{y}$ in $\tilde{x}_{1:n}$. Since $\tilde{x}_{1:n}$ is a uniformly random permutation of $x_{1:n}$, itself drawn from a memoryless source, we can use the same argument as in Lemma 4 to view $\tilde{x}_{1:n}$ as drawn from a random process which outputs $\tilde{x}_1, \tilde{x}_2, \ldots$ according to μ. The times of first occurrence of symbols in $\mathbf{y}$ are thus

$$\begin{aligned}
\tau_t(y_1) &= 1 \\
\tau_t(y_2) &\sim 1 + \mathscr{G}(1 - \mu(y_1)) = \tau_t(y_1) \\
&\vdots \\
\tau_t(y_i) &\sim \tau_t(y_{i-1}) + \mathscr{G}\left(1 - \sum_{j=1}^{i-1} \mu(y_j)\right),
\end{aligned}$$

approximately (i.e. when t is large enough); observe that this approximation is an upper bound on $\tau_t(y_i)$, since at most t symbols can be observed. Since $D_{t,i} = \tau_n(y_i) - 1$ for $i \leq K$, we deduce that

$$\begin{aligned}
D_{t,K+1} &\sim \sum_{i=1}^{K} \mathscr{G}\left(1 - \sum_{j=1}^{i-1} \mu(y_j)\right) - 1, \text{ and} \\
\mathbf{E}[D_{t,K+1}] &\leq K\,\mathbf{E}\left[\mathscr{G}\left(1 - \sum_{j=1}^{K} \mu(y_j)\right)\right] - 1 \\
&\leq K(1 - \mu(\mathcal{Z}_{\text{AUG}}))^{-1} - 1,
\end{aligned}$$

as required. □

Of course, since $D_{t,K+1} < t$, this upper bound is trivial for distributions such as the problematic $\mu(\cdot) \propto 2^i$ described in Section 6 of the main text.

Lemma 7. *Let $\mathscr{U}(\cdot)$ denote the discrete uniform distribution and let $P(\cdot)$ be an arbitrary distribution over $\mathbb{N}$. For $t \in \mathbb{N}^+$, let*

$$\theta \sim P(\cdot) \qquad r \sim \mathscr{U}(\{1, \ldots, t\}).$$

Then

$$\Pr\{r \geq \theta\} \geq \Pr\{r \geq \mathbf{E}\,\theta\}.$$

Proof.

$$\begin{aligned}
\Pr\{r \geq \theta\} &= \sum_{j=0}^{\infty} \Pr\{\theta = j\} \sum_{i=1}^{t} \Pr\{r = i\} \mathbb{I}_{[i \geq j]} \\
&= \sum_{j=1}^{\infty} \Pr\{\theta = j\} \frac{t - j + 1}{t} \mathbb{I}_{[j \leq t]} + \Pr\{\theta = 0\} \\
&= \frac{1}{t} \sum_{j=1}^{\infty} \Pr\{\theta = j\} \left[(t - j + 1)\mathbb{I}_{[j \leq t]} + t\mathbb{I}_{[j > t]} - t\mathbb{I}_{[j > t]} \right] + \Pr\{\theta = 0\} \\
&= \sum_{j=0}^{\infty} \Pr\{\theta = j\} - \frac{1}{t} \sum_{j=1}^{\infty} \Pr\{\theta = j\} \left[(j - 1)\mathbb{I}_{[j \leq t]} + t\mathbb{I}_{[j > t]} \right] \\
&\geq 1 - \frac{\mathbf{E}\,\theta}{t},
\end{aligned}$$

But then

$$\Pr\{r \geq \mathbf{E}\,\theta\} = 1 - \frac{\lceil \mathbf{E}\,\theta \rceil}{t} \leq 1 - \frac{\mathbf{E}\,\theta}{t},$$

and the result follows. □

Lemma 8. *Let z_t be defined as in Algorithm 2 (Budget SAD), let $\kappa := K(1 - \mu(\mathcal{Z}_{\text{AUG}}))^{-1}$, and assume that $G_n(\mathcal{Z}^*)$ occurs (Lemma 5). Then*

$$(1 - \mu(\mathcal{Z}^*))\,(t - \kappa \log t) \leq \mathbf{E}\,[z_t] \leq (1 - \mu(\mathcal{Z}^*))t$$

$$0 \leq \mathbf{E}\,[z_t]\,.$$

Proof. We begin by writing z_t as a sum of indicator functions:

$$\begin{aligned}
z_t &= \sum_{i=1}^{t} \mathbb{I}_{[x_i \text{ is discarded}]} \\
&= \sum_{i=1}^{t} \mathbb{I}_{\left[x_i \notin \mathcal{Z}^*, r_i > D_{i,K+1}\right]} \qquad (4)
\end{aligned}$$

Observe that $\{x_i \notin \mathcal{Z}^*\}$ is independent from $\{r_i > D_{i,K+1}\}$. Using the fact that μ is memoryless, we further have

$$\mathbf{E}\,[z_t] = (1 - \mu(\mathcal{Z}^*)) \sum_{i=1}^{t} \Pr\{r_i > D_{i,K+1}\}\,,$$

from which we immediately infer the upper bound. Then, since $r_i \sim \mathscr{U}(\{0,\dots,i-1\})$, we appeal to Lemmas 6 and 7 to write

$$\begin{aligned}
\mathbf{E}\left[z_t\right] &\geq (1-\mu(\mathcal{Z}^*))\sum_{i=1}^{t}\Pr\left\{r_i > \mathbf{E}\left[D_{i,K+1}\right]\right\} \\
&= (1-\mu(\mathcal{Z}^*))\sum_{i=1}^{t}\left(1-\frac{\mathbf{E}\left[D_{i,K+1}\right]+1}{i}\right) \\
&\geq (1-\mu(\mathcal{Z}^*))\sum_{i=1}^{t}\left(1-\frac{K(1-\mu(\mathcal{Z}_{\text{AUG}}))^{-1}}{i}\right) \\
&\geq (1-\mu(\mathcal{Z}^*))\left(t-\left(K(1-\mu(\mathcal{Z}_{\text{AUG}}))^{-1}\right)\log t\right) \\
&= (1-\mu(\mathcal{Z}^*))\left(t-\kappa\log t\right),
\end{aligned}$$

yielding the lower bound, as desired. □

Theorem 2. *Let $\delta = o\left(\log|\mathcal{X}| + n\log n\right)$ and $K \geq \upsilon(\mathcal{Z}^*,\mu)^{-1}\log(|\mathcal{Z}^*|n\delta^{-1})$. Let $\mathcal{Z}_{\text{AUG}} \subseteq \mathcal{X}$ of size K maximizing $\mu(\cdot)$, and let $\kappa := K(1-\mu(\mathcal{Z}_{\text{AUG}}))^{-1}$. The expected redundancy of the Budget SAD ρ^{B} with respect to the optimal ρ^*_μ is bounded as*

$$\mathbf{E}\left[\mathfrak{F}_n(\rho^{\text{B}},\rho^*_\mu)\right] \leq \mu(\mathcal{Z}^*)\tfrac{|\mathcal{Z}^*|-1}{2}\log n + (1-\mu(\mathcal{Z}^*))\kappa\log^2 n + \ O(\kappa\log(\kappa n)\log\log n).$$

Proof. Let $\rho^{\text{B}}_t(\cdot) := \rho^{\text{B}}(\cdot\,\|\,x_{<t})$, $G_n := G_n(\mathcal{Z}^*)$, $\gamma_t := \gamma(t,\mathcal{Y}_t)$, and let $\mathcal{A}_t := \{x\in\mathcal{X} : N_t(x) > 0\}$ be the true observed alphabet. Recall that for $x\in\mathcal{Y}_t$,

$$\rho^{\text{B}}_t(x) \propto \tilde{N}_t(x) = \max\left\{\hat{N}_t(x), \hat{\gamma}_t w_t(x)\right\},$$

and also

$$\sum_{x\in\mathcal{X}}\rho^{\text{B}}_t(x) = \sum_{x\in\mathcal{Y}_t}\tilde{N}_t(x) + \hat{\gamma}_t w_t(x).$$

In what follows we assume without loss of generality that this sum is equal to

$$\sum_{x\in\mathcal{Y}_t}\hat{N}_t(x) + \hat{\gamma}_t w_t(x) = t + \gamma(t,\mathcal{Y}_t),$$

noting that even when this assumption does not hold the excess redundancy (under the other conditions of the theorem) can be shown to be negligible. We begin by expanding the definition of expected regret:

$$\begin{aligned}
\mathop{\mathbf{E}}_{x_{1:n},r_{1:n}}\left[\mathfrak{F}_n(\rho^{\text{B}},\rho^*_\mu)\right] &= \mathop{\mathbf{E}}_{x_{1:n},r_{1:n}}\left[-\log\frac{\rho^{\text{B}}(x_{1:n})}{\rho^*_\mu(x_{1:n})}\right] \\
&= \mathop{\mathbf{E}}_{x_{1:n},r_{1:n}}\left[-\sum_{t=1}^{n}\log\frac{\rho^{\text{B}}_t(x_t)}{\rho^*_\mu(x_t)}\right] \\
&= \sum_{t=1}^{n}\mathop{\mathbf{E}}_{x_{1:n},r_{1:n}}\left[-\log\frac{\rho^{\text{B}}_t(x_t)}{\rho^*_\mu(x_t)}\right] \\
&= \sum_{t=1}^{n}\mathop{\mathbf{E}}_{x_{1:t},r_{1:t}}\left[-\log\frac{\rho^{\text{B}}_t(x_t)}{\rho^*_\mu(x_t)}\right]. \qquad (5)
\end{aligned}$$

For each time $t \in [n]$ we consider three disjoint events: 1) $\{x_t \in \mathcal{Z}^*\} \cap G_n$, 2) $\{x_t \in \mathcal{Z}^*\} \cap \neg G_n$, and 3) $x_t \notin \mathcal{Z}^*$. We write

$$\mathop{\mathbf{E}}_{x_{1:t}, r_{1:t}} \left[-\log \frac{\rho_t^{\mathrm{B}}(x_t)}{\rho_\mu^*(x_t)} \right] = \mathop{\mathbf{E}}_{x_{1:t}, r_{1:t}} \left[-\log \frac{\rho_t^{\mathrm{B}}(x_t)}{\rho_\mu^*(x_t)} \left(\mathbb{I}_{[x_t \in \mathcal{Z}^* \cap G_n]} + \mathbb{I}_{[x_t \in \mathcal{Z}^* \cap \neg G_n]} + \mathbb{I}_{[x_t \notin \mathcal{Z}^*]} \right) \right]$$

We now bound each term individually.

Case 1. $\{x_t \in \mathcal{Z}^*\} \cap G_n$. By definition of G_n, we have that $c_{t,I(x)} = N_t(x)$ for all $x \in \mathcal{Z}^*$. Furthermore, $(t - z_t) \geq 1$. The Budget SAD predicts x_t with probability

$$\begin{aligned} \rho_t^{\mathrm{B}}(x_t) &= \frac{\tilde{N}_t(x)}{\sum_{x \in \mathcal{Y}_t} \tilde{N}_t(x) + \hat{\gamma}_t} \\ &\geq \frac{\hat{N}_t(x)}{t + \gamma(t, \mathcal{Y}_t)} \\ &\geq \frac{N_t(x)}{t + \gamma(t, \mathcal{A}_t)} = \rho^{\mathrm{SAD}}(x_t \,|\, x_{<t}). \end{aligned}$$

Furthermore, for any $x \in \mathcal{Z}^*$, $\rho_\mu^*(x) = \mu(x)$. Therefore, if $x_t \in \mathcal{Z}^*$ and $G_n(\mathcal{Z}^*)$, then

$$-\log \frac{\rho_t^{\mathrm{B}}(x_t)}{\rho_\mu^*(x_t)} \leq -\log \frac{\rho^{\mathrm{SAD}}(x_t \,|\, x_{<t})}{\mu(x_t)}.$$

In other words, given G_n, the Budget SAD predicts better on $x_t \in \mathcal{Z}^*$ than the SAD.

Case 2. $\{x_t \in \mathcal{Z}^*\} \cap \neg G_n$. Recall that, by definition, we have

$$\rho_t^{\mathrm{B}}(x_t) \geq \frac{w_t(x_t)(z_t + \gamma_t)}{t + \gamma_t}. \tag{6}$$

We thus conservatively bound the loss of the Budget SAD as

$$\begin{aligned} \mathop{\mathbf{E}}_{x_{1:t}, r_{1:t}} \left[-\log \frac{\rho_t^{\mathrm{B}}(x_t)}{\rho_\mu^*(x_t)} \,\middle|\, x_t \in \mathcal{Z}^*, \neg G_n \right] &\leq -\log \frac{w_t(x_t)(z_t + \gamma_t)}{t + \gamma_t} + \log \mu(x_t) \\ &\leq \log |\mathcal{X} \setminus \mathcal{Y}_t| + \log \frac{t + \gamma_t}{z_t + \gamma_t} \\ &\leq \log |\mathcal{X}| + \log \frac{t + \gamma_t}{\gamma_t} \\ &\leq \log |\mathcal{X}| + \log t - \log \gamma_t + 1 \\ &\leq \log |\mathcal{X}| + \log t + \log \log t + 1, \end{aligned} \tag{7}$$

where we assumed that $t \geq \gamma_t$ without loss of generality.

Case 3. $x_t \notin \mathcal{Z}^*$.

Let

$$f(t) := \max \{t - \kappa \log t, 0\}.$$

From Equation 6 and Lemma 8, we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{x_{1:t},r_{1:t}}\left[-\log\frac{\rho_t^{\mathrm{B}}(x_t)}{\rho_\mu^*(x_t)}\,\middle|\,x_t\notin\mathcal{Z}^*\right] &\leq -\log\left[\frac{w_t(x_t)(z_t+\gamma_t)}{t+\gamma_t}\frac{|\mathcal{X}\setminus\mathcal{Z}^*|}{1-\mu(\mathcal{Z}^*)}\right]\\
&\leq -\log\left[\frac{|\mathcal{X}\setminus\mathcal{Z}^*|}{|\mathcal{X}\setminus\mathcal{Y}_t|}\frac{(1-\mu(\mathcal{Z}^*))f(t)+\gamma_t}{t+\gamma_t}\left(1-\mu(\mathcal{Z}^*)\right)^{-1}\right]\\
&\leq -\log\left[\max\left\{\frac{(1-\mu(\mathcal{Z}^*))f(t)}{t},\frac{\gamma_t}{t+\gamma_t}\right\}\left(1-\mu(\mathcal{Z}^*)\right)^{-1}\right]\\
&\leq \min\left\{-\log\left[\frac{f(t)}{t}\right],-\log\left[\frac{(1-\mu(\mathcal{Z}^*))^{-1}\gamma_t}{t+\gamma_t}\right]\right\}\\
&\leq \min\left\{-\log\left[\frac{f(t)}{t}\right],O(\log t)\right\}. \qquad (8)
\end{aligned}$$

We now sum over all time steps. For the event $\{x_t\in\mathcal{Z}\}\cap G_n$, we use the SAD estimation bound from Hutter (2013; see also the main document):

$$\sum_{t=1}^{n}\mathop{\mathbf{E}}_{x_{1:t},r_{1:t}}\left[-\log\frac{\rho_t^{\mathrm{B}}(x_t)}{\rho_\mu^*(x_t)}\,\middle|\,x_t\in\mathcal{Z},G_n\right]\leq\frac{|\mathcal{Z}|-1}{2}\log n+O\left(\log\log n\right). \qquad (9)$$

Combining $T:=\lceil\kappa\log n\rceil$ with Equation 8 yields

$$\begin{aligned}
\sum_{t=1}^{n}\mathop{\mathbf{E}}_{x_{1:t},r_{1:t}}\left[-\log\frac{\rho_t^{\mathrm{B}}(x_t)}{\rho_\mu^*(x_t)}\,\middle|\,x_t\notin\mathcal{Z}^*\right] &\leq O(T\log T)-\sum_{t=T+1}^{n}\log\left[\frac{t-\kappa\log t}{t}\right]\\
&\leq O(T\log T)+\sum_{t=T+1}^{n}\frac{\kappa\log t}{t-\kappa\log t}\\
&\leq O(T\log T)+\kappa\log^2 n \qquad (10)
\end{aligned}$$

Combining Equations 7, 9, and 10, we write

$$\begin{aligned}
\mathop{\mathbf{E}}_{x_{1:n},r_{1:n}}\left[\mathfrak{F}_n(\rho^{\mathrm{B}},\rho_\mu^*)\right] &\leq \Pr\{x_t\in\mathcal{Z}^*\cap\neg G_n\}\left(n\log|\mathcal{X}|+2n\log n\right)\\
&\quad+\Pr\{x_t\in\mathcal{Z}^*\cap G_n\}\left(\frac{|\mathcal{Z}^*|-1}{2}\log n+O\left(\log\log n\right)\right)\\
&\quad+\Pr\{x_t\notin\mathcal{Z}^*\}\left(O(T\log T)+\kappa\log^2 n\right)\\
&= \delta\mu(\mathcal{Z}^*)\left(n\log|\mathcal{X}|+2n\log n\right)\\
&\quad+(1-\delta)\mu(\mathcal{Z}^*)\left(\frac{|\mathcal{Z}^*|-1}{2}\log n+O(\log\log n)\right)\\
&\quad+(1-\mu(\mathcal{Z}^*))(\kappa\log^2 n+O(T\log T)), \qquad (11)
\end{aligned}$$

and taking $\delta=o(n(\log n+\log|\mathcal{X}|))$ we have

$$\mathop{\mathbf{E}}_{x_{1:n},r_{1:n}}\left[\mathfrak{F}_n(\rho^{\mathrm{B}},\rho_\mu^*)\right]\leq\mu(\mathcal{Z}^*)\frac{|\mathcal{Z}^*|-1}{2}\log n+(1-\mu(\mathcal{Z}^*))\kappa\log^2 n+O(\kappa\log(\kappa n)\log\log n)$$

since $T=O(\kappa\log n)$.

As a final remark, note that we can alternatively bound the left hand side of Equation 10 by $O(n\log n)$, and therefore also bound the whole loss by $O(n\log n)$. □

References

Cover, T. M. and Thomas, J. A. (1991). *Elements of information theory*. John Wiley & Sons.

Hutter, M. (2013). Sparse adaptive dirichlet-multinomial-like processes. In *COLT*.