Online Probability Model Estimation For Video Compression

Yue Sun †‡  Jingning Han‡  Yaowu Xu‡

† University of Washington, Seattle  ‡ Google Inc.

March 2020
Probability estimation in video coding system

The minimum codelength coincides with the entropy

\[ H(s_1, s_2, \ldots, s_t) = - \sum_{i=1}^{t} p^*(s_i | s_{i-1} \ldots s_1) \log_2 p^*(s_i | s_{i-1} \ldots s_1). \]

A good estimation will give us a small code length.
Baselines

• Suppose \( p^* (s_t|s_{t-1}...s_1) = p^*(s_i) \), i.e., the symbols are i.i.d. At time \( t \), symbol 0 appears \( k \) times, then the estimation\(^1\)
  \[ p(0) = \frac{k+1}{t+2}. \]

• CABAC\(^2\) and AV1\(^3\) use the update rule
  \[
  p(1)_t = ap(1)_{t-1} + (1 - a)s_t \\
  = (1 - a)(s_t + as_{t-1} + a^2s_{t-2} + ... + a^{t-1}s_1) + a^t p(1)_0.
  \]

\(^1\)Suppose the prior of 0 and 1 are both 0.5.
\(^2\)Wiegand et al., Overview of the h. 264/avc video coding standard.
\(^3\)Chen et al., An overview of core coding tools in the av1 video codec.
Principle of new algorithms

2. Robustness to noise.
3. Computational efficiency for real time application.
4. * Adaptivity with respect to data.

Due to the interest of 2. and 3., we do not use complex models such as neural networks or combinatorial methods.
Warmup: Second order system$^4$

If the update rule is

\[
q^+ = aq + (1 - a)u, \\
r^+ = br + (1 - b)u, \\
p = wq + (1 - w)r, \quad w \in (0, 1).
\]

then the probability estimation update is the following second order linear system

\[
 p_{t+1} = (a + b)p_t - abp_{t-1} \\
+ (w(1 - a) + (1 - w)(1 - b))u_t + (ab - (1 - w)a - wb)u_{t-1}.
\]

---

$^4$Alshin et al., High precision probability estimation for cabac,” in 2013 Visual Communications and Image Processing (VCIP).
Idea: Fixed or adaptive aggregation of algorithms

We can average more than 2 update rules.

- Many algorithms work, for example, \( p_+ = a p + (1 - a) u \) with different \( a \) all roughly do the job.
- Denote \( p \in \mathbb{R}^{np \times 2} \), each row of \( p \) corresponds to a “good” update rule. Denote coefficients \( w \in \mathbb{R}^{np} \), such that \( w \geq 0 \), \( \mathbf{1}^T w = 1 \).
- Then we use \( w^T p \) – weighted average of baseline estimations as final estimation.
Idea: Fixed or adaptive aggregation of algorithms

We can average more than 2 update rules.

\[ p_i^+ = a p_i + (1 - a) u, \]
\[ p = \sum_i w_i p_i. \]

The problem is that we do not know \( w \). We can either pick a fixed \( w \), or even update \( w \) online.
Adaptive aggregation of algorithms

Fixed algorithms

AV1

Counting

\[ p_+ = 0.99p + 0.01s \]

\[ p_+ = 0.98p + 0.02s \]

......

\[ p_+ = 0.8p + 0.2s \]
Adaptive aggregation of algorithms

For each symbol $s_t$, we incur the entropy

$$f(w, p; s_t) = -\log_2((w^T p)(s_t)),$$

and we take gradient with respect to $w$,

$$\nabla_w f(w, p; s_t) = -\frac{1}{(w^T p)(s_t)} p(:, s_t).$$

At each time, we run a gradient step\(^5\)

$$w \leftarrow \arg\min_{w_+ > 0, 1^T w_+ = 1} \| w_+ - w \|^2_2 + 2\eta_t (w_+ - w)^T \nabla_w f(w, p; s_t)$$

$$= \arg\min_{w_+ > 0, 1^T w_+ = 1} \| w_+ - (w - \eta_t \nabla_w f(w, p; s_t)) \|^2.$$

$$\eta_t = 1, 1/2, 1/3, 1/4, 1/5, 1/6, 1/7, \ldots$$

$$\eta_t = 1, 1/4, 1/4, 1/4, 1/4, 1/4, 1/9, 1/9, \ldots$$

\(^5\)We can do a batch algorithm, where we take average of gradients with batch size $1, 4, 9, \ldots$ and update with fixed step size at the end of batches. $O(t^{1/3})$ updates until time $t$. 
Fast projection

The dual of

$$\min_{x \geq 0, 1^T x = 1} \frac{1}{2} \|x - y\|^2.$$ 

is

$$\max_{\mu} \frac{1}{2} \| \max(y - \mu 1, 0) - y \|^2 + \mu (1^T \max(y - \mu 1, 0) - 1),$$

and correspondingly,

$$x_i = \max(y_i - \mu, 0).$$

1 dimensional convex optimization, solve by binary search. To reach $\epsilon$ accuracy for $\mu^*$, need $-\log_2(\epsilon)$ function evaluations.
Experiments

Figure 1: Codeword length reduction on synthetic data. Generate symbols with Bernoulli 0.01, 0.3 alternatively with chunk size 50, 100, 200, 1000. We plot the improvement of entropy of the proposed Multimodal SGD as compared to CABAC.
## Experiments

<table>
<thead>
<tr>
<th>algorithm/dataset</th>
<th>200</th>
<th>400</th>
<th>800</th>
<th>1200</th>
<th>2000</th>
<th>2800</th>
<th>3600</th>
<th>5200</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multimodal Fixed</td>
<td>1368</td>
<td>2587</td>
<td>2940</td>
<td>3860</td>
<td>3770</td>
<td>3709</td>
<td>3465</td>
<td>2388</td>
</tr>
<tr>
<td>Multimodal SGD</td>
<td>1364</td>
<td>2571</td>
<td>2930</td>
<td>3822</td>
<td>3733</td>
<td>3671</td>
<td>3433</td>
<td>2363</td>
</tr>
<tr>
<td>Multimodal Batch</td>
<td>1375</td>
<td>2577</td>
<td>2930</td>
<td>3827</td>
<td>3734</td>
<td>3673</td>
<td>3437</td>
<td>2358</td>
</tr>
<tr>
<td>CABAC</td>
<td>1375</td>
<td>2592</td>
<td>2951</td>
<td>3873</td>
<td>3789</td>
<td>3727</td>
<td>3476</td>
<td>2401</td>
</tr>
<tr>
<td>AV1</td>
<td>1382</td>
<td>2580</td>
<td>2939</td>
<td>3843</td>
<td>3760</td>
<td>3698</td>
<td>3455</td>
<td>2380</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>algorithm/dataset</th>
<th>200</th>
<th>400</th>
<th>800</th>
<th>1200</th>
<th>2000</th>
<th>2800</th>
<th>3600</th>
<th>5200</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multimodal Fixed</td>
<td>215</td>
<td>267</td>
<td>477</td>
<td>726</td>
<td>623</td>
<td>472</td>
<td>322</td>
<td>184</td>
</tr>
<tr>
<td>Multimodal SGD</td>
<td>214</td>
<td>265</td>
<td>474</td>
<td>719</td>
<td>613</td>
<td>468</td>
<td>320</td>
<td>185</td>
</tr>
<tr>
<td>Multimodal Batch</td>
<td>213</td>
<td>265</td>
<td>470</td>
<td>717</td>
<td>612</td>
<td>469</td>
<td>319</td>
<td>186</td>
</tr>
<tr>
<td>CABAC</td>
<td>217</td>
<td>269</td>
<td>481</td>
<td>732</td>
<td>627</td>
<td>473</td>
<td>323</td>
<td>186</td>
</tr>
<tr>
<td>AV1</td>
<td>215</td>
<td>267</td>
<td>475</td>
<td>726</td>
<td>619</td>
<td>472</td>
<td>321</td>
<td>184</td>
</tr>
</tbody>
</table>

**Figure 2:** Cheer and harbour test clips in CIF. Entropy of probability estimation algorithms.
Thank you for listening!

Yue Sun

Jingning Han

Yaowu Xu

Thank you for listening!