Optimization-Theoretic Analysis of Nanopore Sequencing Channels

Optimization-Theoretic Analysis of Nanopore Sequencing Channels

Giorgio Taricco
Politecnico di Torino (DET), Turin, I-10129, Italy
Email: taricco@polito.it ORCID: 0000-0003-1981-7494

***Abstract*—Nanopore sequencing produces variable-length, context-dependent observations affected by substitutions, insertions, deletions, and enzyme-driven dwell-time fluctuations. The resulting channel couples symbol uncertainty with synchronization variability and biophysical context. We develop an optimization-theoretic framework for context-dependent insertion–deletion–substitution (CIDS) models augmented with random dwell times. Capacity approximation, parameter estimation, decoding, and redundancy allocation are cast as structured optimization problems via variational mutual-information forms, saddle-point bounds, EM-type likelihood maximization with latent alignments, and constrained redundancy division. Synthetic examples illustrate the proposed bounds and optimization rules. The contribution is primarily theoretical and intended to support future nanopore model calibration and code design.**

I. Introduction

Nanopore sequencing reads DNA via ionic-current fluctuations as single-stranded DNA translocates through a nanopore. Unlike sequencing-by-synthesis platforms that produce discrete observations, nanopore devices output a continuous, noisy current trace whose magnitude depends on multiple nucleotides simultaneously occupying the sensing region. This physical coupling between adjacent bases introduces strong context dependence at the signal level.

Neural base-callers map the resulting time-warped current trace into a variable-length nucleotide sequence. During this process, substitutions arise from noisy signal discrimination, while synchronization errors in the form of insertions and deletions emerge from segmentation uncertainty, dwell-time variability, and alignment decisions. Homopolymers, in particular, are known to induce characteristic deletion patterns, whereas rapid or irregular translocation can generate insertion bursts. These effects violate the assumptions of classical discrete memoryless channels and motivate channel models with memory and synchronization uncertainty.

While classical deletion channels capture certain aspects of synchronization noise, nanopore sequencing exhibits richer structure due to context-dependent substitutions and biophysically induced timing variability. Related challenges arise in *DNA storage systems*, where substitution and synchronization errors must be handled jointly. Existing nanopore abstractions include context-dependent deletion–like models and empirical error characterizations, but a unified theoretical treatment remains limited.

A common context-dependent insertion–deletion–substitution (CIDS) channel with latent alignments and dwell-time variability is used throughout to induce a family of structured optimization problems. Within this framework, capacity approximation arises as variational mutual-information maximization, parameter estimation as likelihood maximization with latent variables via Expectation–Maximization (EM), decoding as shortest-path or dynamic programming optimization, and redundancy allocation as constrained convex optimization. This paper is deliberately theoretical in scope. Rather than modeling a specific nanopore platform, we develop a unified optimization-based framework linking capacity approximation, parameter estimation, decoding, and redundancy allocation under a common CIDS model. The synthetic examples serve only to illustrate the analytical constructions and their qualitative behavior.

A. Related Work

Synchronization channels, especially deletion channels, are analytically difficult; Mitzenmacher's survey [1] reviews bounds, reconstruction, coding methods, and open capacity problems. These difficulties are amplified in nanopore sequencing by context dependence and timing variability.

Information-theoretic studies of DNA storage quantify limits under joint substitution and synchronization noise [2], [3]. Early nanopore experiments established the physical origins of context dependence, dwell-time variability, and alignment-based decoding. In particular, Laszlo *et al.* [4] showed that MspA current levels depend on roughly four nucleotides and used dynamic-programming alignment of current levels to reference genomes, revealing indel-like effects from stochastic translocation. Later nanopore abstractions [5] and neural base-callers surveyed in [6] build on these observations.

Recent work on synchronization channels includes the small-insertion capacity expansion of Tegin and Duman [7] for the binary i.i.d. insertion channel, which exhibits an intrinsic $\alpha \log \alpha$ penalty as $\alpha \to 0$ (α is the insertion probability). Although nanopore channels are richer, this regime provides useful intuition for the asymptotic capacity approximations considered here. A complementary signal-level analysis under geometric dwell-time duplication was developed in [8], where averaging many reads yields an effective linear channel whose spectral structure is characterized through discrete Legendre polynomials.

II. Channel Model

Throughout the paper, the CIDS channel defines the exact probabilistic model, while capacity bounds, dwell-time handling, and auxiliary channels are used as controlled approximations for analysis and computation. This work focuses on a context-dependent insertion–deletion–substitution (CIDS) model that captures the dominant stochastic mechanisms of nanopore sequencing while remaining analytically tractable. Let $\mathcal{B} = \{A, C, G, T\}$ denote the nucleotide alphabet and let $X_1^n = (X_1, \ldots, X_n)$ be the true DNA sequence, where $X_i \in \mathcal{B}$ is the nucleotide at logical sequence position i. The observed output is a random-length sequence Y_1^m, where m varies due to insertion and deletion events.

A. Alignments and Notation

An *alignment* A is a partial one-to-one, order-preserving correspondence between input and output indices [1]. For given lengths (n, m), $A \subseteq \{1, \ldots, n\} \times \{1, \ldots, m\}$. If $(t, s) \in A$, then $(t, s') \notin A$ for all $s' \neq s$ and $(t', s) \notin A$ for all $t' \neq t$ (i.e., each input index t and output index s appears in at most one pair). If $(t, s), (t', s') \in A$ with $t < t'$, then $s < s'$. We denote by $\mathcal{A}(n, m)$ the set of all feasible alignments. Any input index t not appearing in A is a deletion, and any output index s not appearing in A is an insertion. Three components govern the input–output mapping: $i)$ context-dependent substitutions, $ii)$ insertions and deletions, and $iii)$ dwell-time variability. We first formalize the signal-level model underlying nanopore sequencing.

B. Signal Model and Dwell Time

Nanopore ionic current is sampled in time and depends on the nucleotide context occupying the sensing region. The effective sensing region spans k consecutive nucleotides, referred to as a k-mer. The random dwell time $K_i \geq 1$ denotes the number of consecutive current samples generated while the nucleotide at position i occupies the pore. Define

$$S_i \triangleq 1 + \sum_{1 \leq j < i} K_j, \qquad \pi(t) \triangleq \max\{i : S_i \leq t\},$$

so that $\pi(t)$ maps each time index to the corresponding sequence position. Conditioned on the nucleotide context occupying the pore, the ionic current sample at time t is modeled as [6]

$$I_t = h\big(X_{\pi(t):\pi(t)+k-1}\big) + W_t, \tag{1}$$

where $h : \mathcal{B}^k \to \mathbb{R}$ maps each k-mer to its mean ionic current level and W_t accounts for biochemical fluctuations and electronic measurement noise. Dwell-time variability thus induces a random time-warping of the signal without altering the instantaneous emission law.

C. Context-Dependent Substitution Mechanism

Since ionic current depends on local nucleotide context, substitution probabilities are modeled as context dependent. For simplicity, we adopt a first-order context model, which captures the dominant context dependence observed in nanopore base-calling errors while remaining analytically tractable [5], [6], [9]. Let $\theta_{c,a,b}$ denote the probability of observing symbol b when the true local context is (c, a) and observed position s aligns to true index t:

$$P(Y_s = b \mid X_{t-1} = c, X_t = a, (t, s) \in A) = \theta_{c,a,b}, \tag{2}$$

subject to $\theta_{c,a,b} \geq 0$ and $\sum_{b \in \mathcal{B}} \theta_{c,a,b} = 1$.

D. Insertion and Deletion Processes

Insertions and deletions are modeled via a stochastic walk on the alignment grid. An alignment A is generated by a path starting at $(0, 0)$ and ending at (n, m) using only the three elementary moves

- *Match*: $(t, s) \to (t + 1, s + 1)$,
- *Deletion*: $(t, s) \to (t + 1, s)$,
- *Insertion*: $(t, s) \to (t, s + 1)$.

Such paths are order preserving and therefore induce alignments satisfying the monotonicity constraint

$$(t, s), (t', s') \in A, \ t < t' \implies s < s'.$$

For a feasible alignment $A \in \mathcal{A}(n, m)$, define the following *projections*: $\pi_X(A) \triangleq \{t \colon \exists\, s \text{ such that } (t, s) \in A\}$ and $\pi_Y(A) \triangleq \{s \colon \exists\, t \text{ such that } (t, s) \in A\}$. Indices $t \notin \pi_X(A)$ correspond to deletions, while indices $s \notin \pi_Y(A)$ correspond to insertions. For an input symbol $X_t = a$, define the step probabilities $p_M(a) = 1 - \delta_a - p_I$, $p_D(a) = \delta_a$, and $p_I \in (0, 1)$. Here, $\delta_a \in [0, 1)$ is the base-dependent deletion probability and p_I is a global insertion probability, chosen so that $p_M(a) \geq 0$ for all $a \in \mathcal{B}$. For a feasible alignment $A \in \mathcal{A}(n, m)$, define the alignment weight

$$w(A \mid X_1^n) \triangleq \prod_{(t,s) \in A} p_M(X_t) \prod_{t \notin \pi_X(A)} p_D(X_t) \prod_{s \notin \pi_Y(A)} p_I.$$

This alignment weight yields the conditional distribution

$$P(A \mid X_1^n, m) = \frac{w(A \mid X_1^n)}{\sum_{A' \in \mathcal{A}(n,m)} w(A' \mid X_1^n)}. \tag{3}$$

We will assume that the output length m is fixed and the normalization is taken over $\mathcal{A}(n, m)$. Insertion symbols are generated independently according to a distribution q_b over $\mathcal{B}$ and are accounted for in $P(Y \mid X, A)$.

E. Full Channel Likelihood

Given input X_1^n, output Y_1^m, and alignment $A \in \mathcal{A}(n, m)$, the joint likelihood is

$$P(Y_1^m, A \mid X_1^n) = P(A \mid X_1^n, m) \prod_{(t,s) \in A} \theta_{X_{t-1}, X_t, Y_s}\, e^{-\gamma(t,s)} \times \prod_{s \notin \pi_Y(A)} q_{Y_s}. \tag{4}$$

where $\gamma(t, s)$ is a dwell-induced alignment cost that accounts for the mismatch between the dwell-time process and the alignment (t, s). Marginalizing over all feasible alignments yields the channel law

$$P(Y_1^m \mid X_1^n) = \sum_{A \in \mathcal{A}(n,m)} P(Y_1^m, A \mid X_1^n). \tag{5}$$

III. Capacity Bounds

Let $X_1^n \sim P_{X_1^n}$ and let Y_1^m denote the corresponding output sequence. The mutual information is

$$I(X_1^n; Y_1^m) = \mathbb{E}\left[\log \frac{P(Y_1^m \mid X_1^n)}{P(Y_1^m)}\right], \tag{6}$$

and the capacity per input base is

$$C = \sup_{P_{X_1^n}} \liminf_{n\to\infty} \frac{1}{n} I(X_1^n; Y_1^m). \tag{7}$$

Memory and synchronization effects make C difficult to evaluate, so that we resort to auxiliary-channel lower and upper bounds.

A. Auxiliary Memoryless Lower Bound

Define an auxiliary *single-letter* memoryless channel $\tilde{W}$ that approximates the true sequence-level channel induced by the CIDS model. For $a, b \in \mathcal{B}$, let $\tilde{W}_{a,b}$ denote the effective probability that an input symbol a produces an output symbol b, obtained by marginalizing over alignments and over all other symbols in the input and output sequences. For an i.i.d. input distribution P_X on $\mathcal{B}$, the induced mutual information is

$$I_{\tilde{W}}(P_X) = \sum_{a,b} P_X(a)\, \tilde{W}_{a,b} \log \frac{\tilde{W}_{a,b}}{\sum_{a'} P_X(a')\, \tilde{W}_{a',b}}. \tag{8}$$

A computable lower bound on the channel capacity is

$$C_{\mathrm{LB}} = \max_{P_X} I_{\tilde{W}}(P_X). \tag{9}$$

Theorem III.1. *If $\tilde{W}_{a,b} > 0$ for all $a, b \in \mathcal{B}$, then the optimization problem* (9) *is strictly concave and admits a unique maximizer $P_X^\star$.*

Proof sketch. Strict positivity of $\tilde{W}$ implies strict concavity of the mutual-information functional over the probability simplex, while compactness of the simplex guarantees existence and uniqueness of the maximizer. □

The variational characterization of mutual information yields the classical Blahut–Arimoto (BA) updates [10], [11], which follow from the KKT optimality conditions [12]. Starting from an initialization $P_X^{(0)}$ and iterating for $k = 0, 1, 2, \ldots$, the updates are

$$Q_b^{(k)} = \sum_a P_X^{(k)}(a)\, \tilde{W}_{a,b}, \tag{10}$$

$$P_X^{(k+1)}(a) = \frac{\exp\Big(\sum_b \tilde{W}_{a,b} \log \frac{\tilde{W}_{a,b}}{Q_b^{(k)}}\Big)}{\sum_{a'} \exp\Big(\sum_b \tilde{W}_{a',b} \log \frac{\tilde{W}_{a',b}}{Q_b^{(k)}}\Big)}. \tag{11}$$

B. Dual Upper Bound via Auxiliary Channels

Let $Q_{Y|X}$ be any auxiliary channel. By the variational characterization of mutual information underlying the BA algorithm [10], [11],

$$I(X;Y) \le \mathbb{E}\left[\log \frac{Q_{Y|X}(Y|X)}{Q_Y(Y)}\right], \tag{12}$$

with Q_Y induced by $(P_X, Q_{Y|X})$. Minimizing over all channels $Q_{Y|X}$, which includes the true channel $Q_{Y|X} = P_{Y|X}$, yields the exact variational representation

$$C = \min_{Q_{Y|X}} \max_{P_X} \mathbb{E}\left[\log \frac{Q_{Y|X}(Y|X)}{Q_Y(Y)}\right]. \tag{13}$$

Equation (13) motivates auxiliary-channel upper bounds obtained by restricting $Q_{Y|X}$ to a tractable family. This dual viewpoint also yields computable upper proxies when the auxiliary family is restricted to a tractable class, thereby complementing the auxiliary-channel lower bound.

C. Markov Input Generalization

For a stationary first-order Markov input with transition probabilities $T_{a,c} = P(X_t = c \mid X_{t-1} = a)$ and stationary probability distribution π, $I(X;Y) =$

$$\sum_{a,c} \pi(a) T_{a,c} \sum_b \tilde{W}_{c,b} \log \frac{\tilde{W}_{c,b}}{\sum_{a',c'} \pi(a') T_{a',c'} \tilde{W}_{c',b}}. \tag{14}$$

The maximization is over stationary first-order Markov inputs with stationary distribution π and transition matrix T:

$$\pi = \pi T, \qquad \sum_c T_{a,c} = 1, \qquad T_{a,c} \ge 0. \tag{15}$$

Accordingly,

$$C_{\mathrm{LB}}^{(1)} = \max_{\pi, T} I(X;Y). \tag{16}$$

Lemma III.2. *The objective* (14) *is jointly concave in (π, T) over the Markov polytope* (15)*.*

Proof sketch. Mutual information is concave in the input distribution; $(\pi, T) \mapsto p_X$ is linear on the polytope. □

D. Time-Scaled Capacity Approximation

Let $\kappa(a) = \mathbb{E}[K_t \mid X_t = a]$ and $\bar{\kappa} \triangleq \sum_a P_X(a)\kappa(a)$. Then, a throughput-normalized approximation is

$$C_{\mathrm{TS}} = \frac{1}{\bar{\kappa}} \max_{P_X} I_{\tilde{W}}(P_X). \tag{17}$$

This normalization accounts for the average dwell time per input symbol, so that C_{TS} represents information rate per unit time rather than per emitted symbol.

IV. Parameter Estimation

Let Θ collect $\theta_{c,a,b}$, δ_a, q_b, and dwell-time parameters $p_k(a)$. Given training pairs $(X^{(i)}, Y^{(i)})$, the likelihood is

$$\begin{aligned} \mathcal{L}(\Theta) &= \prod_i P_\Theta\big(Y^{(i)} \mid X^{(i)}\big) \\ &= \prod_i \sum_{A \in \mathcal{A}(n_i, m_i)} P_\Theta\big(Y^{(i)}, A \mid X^{(i)}\big), \end{aligned} \tag{18}$$

and the log-likelihood

$$\ell(\Theta) = \sum_i \log\Big(\sum_A P_\Theta(Y^{(i)}, A \mid X^{(i)})\Big). \tag{19}$$

Here, each $X^{(i)}$ denotes a complete input sequence used for training, with $Y^{(i)}$ the corresponding observed output sequence. The latent alignment couples substitutions, insertions, deletions, and dwell effects, so inference reduces to estimating a hidden path on the alignment lattice together with the model parameters.

A. E-step: Alignment Posteriors

Given (X_1^n, Y_1^m) and parameters $\Theta^{(\ell)}$, the E-step computes posterior alignment quantities $P_{\Theta^{(\ell)}}(A \mid X, Y)$ without enumerating $A \in \mathcal{A}(n,m)$. This is achieved via a forward–backward (BCJR) recursion on the alignment lattice, equivalently the E-step of a pair-HMM [13]–[15]. Define forward and backward variables for $t = 0, \ldots, n$, $s = 0, \ldots, m$,

$$\alpha(t,s) \triangleq P(Y_1^s,\ \text{reach } (t,s) \mid X_1^t), \tag{20}$$

$$\beta(t,s) \triangleq P(Y_{s+1}^m,\ \text{go } (t,s) \to (n,m) \mid X_{t+1}^n). \tag{21}$$

Let $Z \triangleq \alpha(n,m)$ denote the unnormalized partition function. Forward recursion:

$$\begin{aligned}\alpha(t,s) = {}& \alpha(t-1,s-1)\, p_M(X_t)\, \theta_{X_{t-1},X_t,Y_s}\, e^{-\gamma(t,s)} \\ &+ \alpha(t-1,s)\, p_D(X_t) + \alpha(t,s-1)\, p_I\, q_{Y_s},\end{aligned} \tag{22}$$

with $\alpha(0,0) = 1$,$\alpha(t,0) = \alpha(t-1,0)p_D(X_t)$, $\alpha(0,s) = \alpha(0,s-1)p_I q_{Y_s}$. Backward recursion:

$$\begin{aligned}\beta(t,s) = {}& p_M(X_{t+1})\, \theta_{X_t,X_{t+1},Y_{s+1}}\, e^{-\gamma(t+1,s+1)} \\ & \beta(t+1,s+1) + p_D(X_{t+1})\, \beta(t+1,s) \\ & + p_I\, q_{Y_{s+1}}\, \beta(t,s+1),\end{aligned} \tag{23}$$

with $\beta(n,m)$ =1, $\beta(t,m) = p_D(X_{t+1})\beta(t+1,m)$, $\beta(n,s) = p_I q_{Y_{s+1}} \beta(n,s+1)$. Match-edge posterior:

$$\begin{aligned}&P(\chi_{t,s} = 1 \mid X, Y) \\ &= \frac{\alpha(t-1,s-1)\, p_M(X_t)\, \theta_{X_{t-1},X_t,Y_s}\, e^{-\gamma(t,s)}\, \beta(t,s)}{Z}.\end{aligned} \tag{24}$$

Deletion and insertion posteriors:

$$P(X_t \text{ deleted} \mid X, Y) = \sum_{s=0}^{m} \frac{\alpha(t-1,s)\, p_D(X_t)\, \beta(t,s)}{Z},$$

$$P(Y_s \text{ inserted} \mid X, Y) = \sum_{t=0}^{n} \frac{\alpha(t,s-1)\, p_I\, q_{Y_s}\, \beta(t,s)}{Z}.$$

For an input of length n and output of length m, the forward–backward recursions require $\mathcal{O}(nm)$ time and $\mathcal{O}(nm)$ memory in the worst case; standard banding or pruning reduces this when posterior mass concentrates near the main diagonal.

B. M-step: Closed-Form Updates

Maximize

$$Q(\Theta \mid \Theta^{(\ell)}) = \mathbb{E}_{A|X,Y,\Theta^{(\ell)}}\big[\log P_\Theta(Y, A \mid X)\big]. \tag{25}$$

Updates are

$$\theta_{c,a,b}^{(\ell+1)} = \frac{N_{c,a,b}}{\sum_{b' \in \mathcal{B}} N_{c,a,b'}},\quad \delta_a^{(\ell+1)} = \frac{D_a}{D_a + \sum_b N_{\star,a,b}}, \tag{26}$$

$$q_b^{(\ell+1)} = \frac{I_b}{\sum_{b'} I_{b'}}, \tag{27}$$

where $N_{\star,a,b} \triangleq \sum_{c \in \mathcal{B}} N_{c,a,b}$. Dwell-time parameters are assumed fixed (or pre-estimated) and are not updated within the EM iterations; the term $\gamma(t,s)$ therefore acts as a fixed alignment cost.

Theorem IV.1. *Under the CIDS model, EM iterates satisfy* $\ell(\Theta^{(\ell+1)}) \geq \ell(\Theta^{(\ell)})$.

Proof sketch. Standard EM monotonicity follows from Jensen's inequality applied to latent alignments. □

C. Variational Inference

For long reads, approximate inference uses a tractable family $q(A)$ [1] and maximizes the evidence lower bound (ELBO)

$$\begin{aligned}\mathcal{F}(q,\Theta) \triangleq {}& \mathbb{E}_{q(A)}[\log P_\Theta(Y, A \mid X)] \\ & - D_{\mathrm{KL}}(q(A) \,\|\, P_\Theta(A \mid X, Y)),\end{aligned} \tag{28}$$

so that $\log P_\Theta(Y|X) \geq \mathcal{F}(q,\Theta)$. Coordinate ascent alternates between optimizing the variational distribution over alignments and the model parameters. Specifically, letting ℓ index the iterations,

$$q^{(\ell+1)} = \arg\max_{q \in \mathcal{Q}} \mathcal{F}\Big(q, \Theta^{(\ell)}\Big), \tag{29}$$

$$\Theta^{(\ell+1)} = \arg\max_{\Theta} \mathcal{F}\Big(q^{(\ell+1)}, \Theta\Big), \tag{30}$$

where $\mathcal{Q}$ denotes a chosen family of tractable distributions over $\mathcal{A}(n,m)$.

V. Decoding and Alignment Optimization

Maximum a posteriori (MAP) decoding seeks

$$\hat{X} = \arg\max_X P(X)\, P(Y \mid X), \tag{31}$$

or, replacing the alignment sum by a maximization,

$$(\hat{X}, \hat{A}) = \arg\max_{X,A} P(X)\, P(A \mid X, m)\, P(Y \mid X, A). \tag{32}$$

Taking negative logarithms of the joint objective in (32), and discarding terms independent of (X, A), define the additive cost

$$\begin{aligned}\Lambda(X,A) \triangleq {}& -\log P(X) - \sum_{(t,s)\in A} \log p_M(X_t) \\ & - \sum_{t \notin \pi_X(A)} \log p_D(X_t) - \sum_{s \notin \pi_Y(A)} \log(p_I q_{Y_s}) \\ & - \sum_{(t,s)\in A} \log \theta_{X_{t-1},X_t,Y_s} + \sum_{(t,s)\in A} \gamma(t,s)\end{aligned} \tag{33}$$

Each term corresponds to a local match, deletion, or insertion event, so minimizing $\Lambda(X,A)$ is equivalent to finding a minimum-cost path on a $(n+1) \times (m+1)$ alignment grid.

[1]Here, $q(A)$ denotes a probability distribution over the set of feasible alignments $\mathcal{A}(n,m)$.

A. Dynamic-Programming Recursion

Let $D(t, s)$ be the minimum cost for prefixes X_1^t and Y_1^s:

$$D(t,s) = \min\Big\{D(t-1,s-1) - \log p_M(X_t) - \log\theta_{X_{t-1},X_t,Y_s} + \gamma(t,s), D(t-1,s) - \log p_D(X_t), D(t,s-1) - \log p_I - \log q_{Y_s}\Big\}. \quad (34)$$

With boundary conditions $D_{0,0} = 0$, $D_{i,0} = i\lambda_D$, and $D_{0,j} = j\lambda_I$, the terminal value $D(n, m)$ and its traceback give the MAP alignment. The recursion has the same $\mathcal{O}(nm)$ worst-case complexity as classical sequence alignment and admits the same banding, pruning, and anti-diagonal parallelization strategies.

Theorem V.1. *Any optimal $(\hat{X}, \hat{A})$ minimizing $\Lambda(X, A)$ corresponds to a minimal-cost path under* (34).

Proof sketch. Optimal substructure implies a Bellman recursion equivalent to (34) [16]. □

B. Normalization and Posterior Decoding

Posterior decoding computes soft alignment quantities like

$$P(\chi_{t,s} = 1 \mid X, Y) = \sum_A \mathbf{1}\{(t,s) \in A\}\, P(A \mid X, Y), \quad (35)$$

via forward–backward recursions over states (t, s), producing soft alignments useful for polishing and error correction.

VI. Coding and Redundancy Optimization

Nanopore codes must address substitutions (via $\theta_{c,a,b}$) and synchronization noise (indels). We cast redundancy division as constrained optimization. Let total redundancy per input base be $R > 0$; allocate r_s to substitution protection and r_i to synchronization, with $r_s + r_i \le R$. Let $P_{\mathrm{e},s}(r_s)$ and $P_{\mathrm{e},i}(r_i)$ be the corresponding error probabilities, so that

$$P_{\mathrm{e}}(r_s, r_i) = P_{\mathrm{e},s}(r_s) + P_{\mathrm{e},i}(r_i) - P_{\mathrm{e},s}(r_s)\, P_{\mathrm{e},i}(r_i). \quad (36)$$

A. Redundancy Allocation Problem

To solve the optimization problem

$$\min_{r_s, r_i \ge 0} P_{\mathrm{e}}(r_s, r_i) \qquad \text{s.t.} \quad r_s + r_i \le R, \quad (37)$$

we build the Lagrangian

$$\mathcal{L}(r_s, r_i, \lambda) = P_{\mathrm{e}}(r_s, r_i) + \lambda(r_s + r_i - R), \quad (38)$$

and get the KKT conditions

$$\frac{\partial P_{\mathrm{e}}}{\partial r_s} + \lambda = 0, \ \frac{\partial P_{\mathrm{e}}}{\partial r_i} + \lambda = 0, \ \lambda(r_s + r_i - R) = 0, \quad (39)$$

where $\lambda \ge 0$. Differentiating (36) gives

$$\frac{\partial P_{\mathrm{e}}}{\partial r_s} = (1 - P_{\mathrm{e},i}(r_i))\, P'_{\mathrm{e},s}(r_s), \quad (40)$$

$$\frac{\partial P_{\mathrm{e}}}{\partial r_i} = (1 - P_{\mathrm{e},s}(r_s))\, P'_{\mathrm{e},i}(r_i), \quad (41)$$

which yield the balance condition

$$(1 - P_{\mathrm{e},i}(r_i))\, P'_{\mathrm{e},s}(r_s) = (1 - P_{\mathrm{e},s}(r_s))\, P'_{\mathrm{e},i}(r_i). \quad (42)$$

Theorem VI.1. *If $P_{\mathrm{e},s}$ and $P_{\mathrm{e},i}$ are strictly decreasing and twice differentiable with $P'_{\mathrm{e},s} < 0$, $P'_{\mathrm{e},i} < 0$, any interior optimum satisfies* (42).

Proof sketch. KKT conditions on $r_s + r_i = R$ yield (42). □

B. Optimization under Exponential Error Approximations

Assume

$$P_{\mathrm{e},s}(r_s) = \exp(-\alpha r_s), \qquad P_{\mathrm{e},i}(r_i) = \exp(-\beta r_i), \quad (43)$$

for $\alpha, \beta > 0$. Then, the solution under $r_s + r_i = R$ is

$$r_s^\star = \frac{R}{2} + \frac{1}{2(\alpha + \beta)} \log\left(\frac{\alpha}{\beta}\right), \qquad r_i^\star = R - r_s^\star. \quad (44)$$

VII. Numerical Example

We use a short synthetic example to illustrate how the lower-bound and redundancy-allocation constructions vary with the channel parameters.

The goal is to demonstrate the analytical framework, not to emulate a calibrated nanopore platform.

Let $\mathcal{B} = \{A, C, G, T\}$ and define the neighbor map $\mathcal{N}$ based on chemical similarity of purines $\{A, G\}$ and pyrimidines $\{C, T\}$. We use a context-independent approximation

$$\theta_{a,b} = \begin{cases} 0.88, & b = a, \\ 0.10, & b = \mathcal{N}(a), \\ 0.01, & \text{otherwise}, \end{cases} \quad (45)$$

so that $\sum_{b \in \mathcal{B}} \theta_{a,b} = 1$ for every $a \in \mathcal{B}$. Let the base-dependent deletion probabilities be

$$\delta_A = 0.03, \quad \delta_C = 0.02, \quad \delta_G = 0.02, \quad \delta_T = 0.04. \quad (46)$$

Define the auxiliary output alphabet $\tilde{\mathcal{Y}} = \mathcal{B} \cup \{\mathsf{e}\}$ where e denotes erasure. Then, the auxiliary DMC $\tilde{W}$ is

$$\tilde{W}(b \mid a) = (1 - \delta_a)\, \theta_{a,b}, \quad b \in \mathcal{B}, \qquad \tilde{W}(\mathsf{e} \mid a) = \delta_a. \quad (47)$$

The BA algorithm yields the lower bound $C_{\mathrm{LB}} = 1.337$.

Now, let the dwell means be

$$\mathbb{E}[K \mid X \in \{A, G\}] = 3.5, \quad \mathbb{E}[K \mid X \in \{C, T\}] = 2.4.$$

The corresponding time-normalized proxy is

$$C_{\mathrm{TS}} = \frac{C_{\mathrm{LB}}}{\bar{\kappa}} = \frac{1.337}{2.950} = 0.453\ \frac{\text{bits}}{\text{current sample}}.$$

Finally, assuming (43) with $\alpha = 5.2$ and $\beta = 4.7$, the KKT conditions (44) yield $r_s^\star = 0.10511$ and $r_i^\star = 0.09489$.

VIII. Conclusion

We proposed a unified optimization-theoretic framework for nanopore sequencing channels based on a context-dependent insertion–deletion–substitution model with dwell-time variability. Within this framework, capacity approximation, parameter learning, decoding, and redundancy allocation become structured optimization problems. Although intentionally stylized and illustrated only by synthetic examples, the framework clarifies the interaction among synchronization uncertainty, context dependence, and dwell effects, and provides a starting point for higher-order models, signal-level likelihoods, and real-data calibration.

References

[1] M. Mitzenmacher, “A survey of results for deletion channels and related synchronization channels,” *Probability Surveys*, vol. 6, pp. 1–33, 2009.

[2] A. Lenz, A. Wachter-Zeh, E. Yaakobi, and J. Bruck, “Coding over sets for dna storage,” *IEEE Transactions on Information Theory*, vol. 66, no. 4, pp. 2331–2351, 2020.

[3] L. Organick, Y.-J. Chen, S. Dumas Ang, R. Lopez, X. Liu, K. Strauss, and L. Ceze, “Probing the physical limits of reliable dna data retrieval,” *Nature Communications*, vol. 11, no. 1, p. 616, 2020.

[4] A. H. Laszlo, I. M. Derrington, B. C. Ross, H. Brinkerhoff, A. Adey, I. C. Nova, J. M. Craig, K. W. Langford, J. M. Samson, R. Daza, K. Doering, J. Shendure, and J. H. Gundlach, “Decoding long nanopore sequencing reads of natural DNA,” *Nature Biotechnology*, vol. 32, no. 8, pp. 829–833, 2014.

[5] W. Mao, S. N. Diggavi, and S. Kannan, “Models and information-theoretic bounds for nanopore sequencing,” *IEEE Transactions on Information Theory*, vol. 64, no. 4, pp. 3216–3237, 2018.

[6] R. R. Wick, L. M. Judd, and K. E. Holt, “Performance of neural network basecalling tools for oxford nanopore sequencing,” *Genome Biology*, vol. 20, no. 1, p. 129, 2019.

[7] B. Tegin and T. M. Duman, “On the capacity of insertion channels for small insertion probabilities,” in *2025 IEEE International Symposium on Information Theory (ISIT)*, 2025, pp. 1–6.

[8] A. Vidal and E. Viterbo, “Spectral analysis of nanopore dna signatures under geometric duplication,” in *2025 IEEE International Symposium on Information Theory (ISIT)*, 2025, pp. 1–6.

[9] F. Rang, W. Kloosterman, and J. de Ridder, “From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy,” *Genome Biology*, vol. 19, no. 1, 13 July 2018.

[10] R. E. Blahut, “Computation of channel capacity and rate-distortion functions,” *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[11] S. Arimoto, “An algorithm for computing the capacity of arbitrary discrete memoryless channels,” *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20, 1972.

[12] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, UK: Cambridge University Press, 2004.

[13] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, “Optimal decoding of linear codes for minimizing symbol error rate (corresp.),” *IEEE Transactions on Information Theory*, vol. 20, no. 2, pp. 284–287, 1974.

[14] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, “A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains,” *The Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164–171, 1970.

[15] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, “Biological sequence analysis: probabilistic models of proteins and nucleic acids,” *Cambridge University Press*, 1998.

[16] R. Bellman, *Dynamic Programming*. Princeton University Press, 1957.