Penn State University
CSE 564 Complexity of Combinatorial Problems

Instructor: Pei Wu

Module

11

Parallel repetition

This lecture explains the *parallel repetition theorem* and how it is used to amplify the soundness gap of two-prover one-round (2P1R) games, equivalently *Label Cover* / projection games. At a very high level, the inapproximability pipeline we have been developing is:

$$\mathrm{NP} \subseteq \mathrm{PCP}[\log n, O(1)] \implies (1, 1/2)\text{-LC is NP-hard} \xrightarrow{\text{parallel repetition}} (1, \delta)\text{-LC is NP-hard}$$

for any constant $\delta > 0$. With this amplified Label Cover gap in hand, one then applies the Long Code and Fourier analysis to obtain hardness of approximation for concrete Boolean CSPs.

We start with one last warm-up reduction (from MAX-3XOR to MAX-3SAT), and then focus on parallel repetition.

11.1 From MAX-3XOR to MAX-3SAT

THEOREM 11.1. *For every $\varepsilon > 0$, it is NP-hard to distinguish between* MAX-3SAT *instances J with*

$$\mathrm{val}(J) \geq 1 - \varepsilon \qquad \textit{and} \qquad \mathrm{val}(J) \leq \frac{7}{8} + \varepsilon.$$

Equivalently, $(1 - \varepsilon, 7/8 + \varepsilon)$-MAX-3SAT *is NP-hard.*

Reduction sketch. We reduce from the gap version of MAX-3XOR established previously: it is NP-hard to distinguish between instances I with $\mathrm{val}(I) \geq 1 - \varepsilon$ and instances with $\mathrm{val}(I) \leq 1/2 + \varepsilon$.

Let I be a MAX-3XOR instance with constraints of the form

$$x_i \oplus x_j \oplus x_k = b, \qquad b \in \{0, 1\}.$$

For each such XOR constraint, we add *four* 3-CNF clauses that jointly enforce the parity check. Concretely, consider first the case $b = 0$. The XOR constraint is *violated* exactly by the four assignments

$$(1,1,1),\ (1,0,0),\ (0,1,0),\ (0,0,1).$$

For each violating assignment (α, β, γ), add a clause that is falsified *only* on that assignment:

$$(\alpha, \beta, \gamma) = (1, 1, 1) \Rightarrow (\neg x_i \vee \neg x_j \vee \neg x_k), \qquad (\alpha, \beta, \gamma) = (1, 0, 0) \Rightarrow (\neg x_i \vee x_j \vee x_k),$$

$$(\alpha, \beta, \gamma) = (0, 1, 0) \Rightarrow (x_i \vee \neg x_j \vee x_k), \qquad (\alpha, \beta, \gamma) = (0, 0, 1) \Rightarrow (x_i \vee x_j \vee \neg x_k).$$

(For $b = 1$, do the analogous construction using the four assignments of *even* parity.)

Let J be the resulting MAX-3SAT instance. Each XOR constraint contributes 4 clauses. A truth assignment that satisfies the XOR constraint satisfies *all* four clauses, whereas an assignment that violates the XOR constraint satisfies *exactly three* of the four clauses. Therefore, for any assignment,

$$\text{fraction of satisfied clauses in } J \;=\; \frac{3}{4} + \frac{1}{4} \cdot \text{fraction of satisfied XOR constraints in } I.$$

In particular,

$$\mathrm{val}(I) \geq 1 - \varepsilon \implies \mathrm{val}(J) \geq 1 - \frac{\varepsilon}{4}, \qquad \mathrm{val}(I) \leq \frac{1}{2} + \varepsilon \implies \mathrm{val}(J) \leq \frac{7}{8} + \frac{\varepsilon}{4}.$$

Renaming $\varepsilon/4$ as ε gives the stated gap. ∎

11.2 2P1R games and Label Cover

Parallel repetition is most naturally stated for 2-prover 1-round games.

DEFINITION 11.2 (2P1R game). A *two-prover one-round game* G consists of:

- finite question sets U, V and answer alphabets Σ_A, Σ_B;
- a distribution μ on pairs of questions $(X, Y) \in U \times V$;
- a verifier predicate
$$\varphi : U \times V \times \Sigma_A \times \Sigma_B \to \{0, 1\}.$$

The referee samples $(X, Y) \sim \mu$, sends X to Alice and Y to Bob, receives answers $a \in \Sigma_A$ and $b \in \Sigma_B$, and accepts iff $\varphi(X, Y, a, b) = 1$.

A (deterministic) strategy is a pair of functions $h_A : U \to \Sigma_A$ and $h_B : V \to \Sigma_B$. The *value* of G is

$$\mathrm{val}(G) \;:=\; \max_{h_A, h_B} \Pr_{(X,Y)\sim\mu} \big[\varphi(X, Y, h_A(X), h_B(Y)) = 1\big].$$

Randomized strategies do not help (by fixing the public coins to the best choice).

A Label Cover instance can be viewed as such a game: the referee samples a random edge (u, v) and asks Alice u and Bob v; the answers are labels, and the predicate checks the projection constraint. Conversely, every 2P1R game can be converted into a constraint satisfaction problem on a bipartite graph. In particular, $\mathrm{val}(I)$ for a Label Cover instance I matches $\mathrm{val}(G)$ for the corresponding game.

11.3 Parallel repetition

Goal. Our goal is to turn a constant-gap Label Cover instance into an arbitrarily small-gap instance. Starting from a constant-query PCP, one can obtain a projection game (Label Cover) with completeness 1 and some constant soundness $s < 1$ (e.g. $s = 1/2$ in the notes). Parallel repetition amplifies soundness.

The repeated game. Let G be a 2P1R game. Its n-fold parallel repetition $G^{\otimes n}$ is the game where the referee samples n independent question pairs

$$(X_1, Y_1), \ldots, (X_n, Y_n) \ \sim\ \mu^{\otimes n},$$

sends $(X_1, \ldots, X_n)$ to Alice and $(Y_1, \ldots, Y_n)$ to Bob, receives answers $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_n)$, and accepts iff

$$\bigwedge_{i=1}^{n} \varphi(X_i, Y_i, a_i, b_i) = 1.$$

THEOREM 11.3 (Parallel repetition (Raz / Holenstein)). *There is an absolute constant $C > 0$ such that for every 2P1R game G and every $n \geq 1$,*

$$\mathrm{val}(G^{\otimes n}) \ \leq\ \Big(1 - \frac{(1 - \mathrm{val}(G))^3}{6000}\Big)^{\frac{n}{\log|\Sigma_A||\Sigma_B|}}.$$

In particular, if $\mathrm{val}(G) \leq 1 - \varepsilon$ *for some constant* $\varepsilon > 0$*, then* $\mathrm{val}(G^{\otimes n})$ *decays exponentially in* $n/\log(|\Sigma_A|\,|\Sigma_B|)$*.*

REMARK 11.4. A tempting (but false) guess would be $\mathrm{val}(G^{\otimes n}) \approx \mathrm{val}(G)^n$. Parallel repetition does give exponential decay, but the exact exponent is subtler because optimal strategies for $G^{\otimes n}$ can correlate answers across coordinates.

As an immediate corollary, if we start with a constant-gap Label Cover instance (equivalently, a game) and repeat it enough times, we obtain $(1, \delta)$-Label Cover hardness for arbitrarily small δ, as desired.

11.4 Proof strategy: conditioning and embedding

We give a proof sketch in the style of the notes (following Holenstein's information-theoretic view).

Fixing a strategy. Fix any (possibly optimal) deterministic strategy for the repeated game $G^{\otimes n}$. Let W_i be the event that the players win in coordinate i:

$$W_i := \big\{\varphi(X_i, Y_i, a_i, b_i) = 1\big\}.$$

By definition of $\mathrm{val}(G)$, for each i we have

$$\mathbb{P}[W_i] \leq \mathrm{val}(G),$$

but the events $W_1, \ldots, W_n$ are *not* independent under an arbitrary strategy.

A typical "one-more-coordinate" lemma. The following is the kind of statement one wants in order to conclude exponential decay.

LEMMA 11.5 ("Conditioning reveals little information" (informal)). *Fix* $\varepsilon > 0$*. Let* $S = \{i_1, \dots, i_t\} \subseteq [n]$ *with* $t \leq \varepsilon n$*, and let*

$$\mathcal{E} := W_{i_1} \wedge \cdots \wedge W_{i_t}.$$

If $\mathbb{P}[E] \geq 2^{-\varepsilon n}$*, then for a typical fresh coordinate* $i \notin S$*,*

$$\mathbb{P}[W_i \mid \mathcal{E}] \;\leq\; \mathrm{val}(G) + \varepsilon.$$

Intuitively, conditioning on E should not make it *much* easier to win in another coordinate, unless E is an extremely rare event. One then iterates Lemma 11.5 to bound

$$\mathbb{P}[W_1 \wedge \cdots \wedge W_n] = \mathrm{val}(G^{\otimes n}).$$

Embedding a single game. A standard way to prove Lemma 11.5 is via an *embedding* argument. Suppose there is a coordinate $i^\star$ such that $\mathbb{P}[W_{i^\star} \mid \mathcal{E}]$ is noticeably larger than $\mathrm{val}(G)$. We will use the repeated-game strategy (together with some sampling tricks) to build a strategy for the *base* game G with success probability larger than $\mathrm{val}(G)$, a contradiction.

Concretely, we try to sample a random "fake transcript"

$$(X_1, \dots, X_n,\ Y_1, \dots, Y_n)$$

distributed *approximately* as in the repeated game conditioned on E. Then we feed $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_n)$ into the fixed repeated-game strategy to obtain answers $(a_1, \dots, a_n)$ and $(b_1, \dots, b_n)$, and we output $(a_{i^\star}, b_{i^\star})$ as our answers for the base game. If the simulated transcript is close enough to the true conditional distribution, we obtain

$$\mathbb{P}[\varphi(X, Y, a_{i^\star}, b_{i^\star}) = 1] \;\approx\; \mathbb{P}[W_{i^\star} \mid \mathcal{E}],$$

violating optimality of $\mathrm{val}(G)$.

The main difficulty is to generate the conditional distribution without allowing Alice and Bob to communicate. This is where *correlated sampling* and a key information-theoretic lemma enter.

11.5 Information theory crash course

All logarithms are base 2.

KL divergence and total variation.

DEFINITION 11.6 (KL divergence). For distributions μ, λ on the same finite space,

$$\mathrm{KL}(\mu \| \lambda) := \mathbb{E}_{x \sim \mu}\left[\log \frac{\mu(x)}{\lambda(x)}\right].$$

Fact 11.7 (Pinsker's inequality).

$$\|\mu - \lambda\|_{\mathrm{TV}}^2 \;\lesssim\; \mathrm{KL}(\mu\|\lambda).$$

(Up to a universal constant factor.)

Fact 11.8 (Total variation distance).

$$\|\mu - \lambda\|_{\mathrm{TV}} \;=\; \max_E \big(\mu(E) - \lambda(E)\big) \;=\; \frac{1}{2}\sum_x |\mu(x) - \lambda(x)|\,.$$

Fact 11.9 (Data processing). *If f is any (possibly randomized) function, then*

$$\mathrm{KL}(f(X)\|f(Y)) \;\le\; \mathrm{KL}(X\|Y), \qquad \|f(X) - f(Y)\|_{\mathrm{TV}} \;\le\; \|X - Y\|_{\mathrm{TV}}\,.$$

Entropy and mutual information. For a random variable X with distribution μ_X,

$$\mathrm{H}(X) := \mathbb{E}\left[\log \frac{1}{\mu_X(X)}\right].$$

For a joint distribution $\mu_{X,Y}$,

$$\mathrm{H}(X \mid Y) := \mathrm{H}(X,Y) - \mathrm{H}(Y), \qquad \mathrm{I}(X;Y) := \mathrm{H}(X) - \mathrm{H}(X \mid Y).$$

Fact 11.10. *Mutual information can be written as an average KL divergence:*

$$\mathrm{I}(X;Y) \;=\; \mathbb{E}_y\Big[\mathrm{KL}\big(\mu_{X|Y=y} \,\|\, \mu_X\big)\Big] \;=\; \mathbb{E}_{x,y}\left[\log \frac{\mu_{X,Y}(x,y)}{\mu_X(x)\mu_Y(y)}\right].$$

In particular $\mathrm{I}(X;Y) \le \mathrm{H}(Y)$.

Conditioning on an event cannot change too many marginals. The following claim appears explicitly in the notes and is a convenient "one-shot" inequality. Let $(X_1,Y_1),\ldots,(X_n,Y_n) \sim \mu^{\otimes n}$ be i.i.d. pairs.

Claim 11.11. *For any event E (measurable w.r.t. the whole transcript),*

$$\sum_{i=1}^n \|(X_i,Y_i) \mid \mathcal{E} - (X_i,Y_i)\|_{\mathrm{TV}}^2 \;\le\; \log \frac{1}{\mathbb{P}[E]}.$$

Proof. By Pinsker (Fact 11.7),

$$\sum_{i=1}^n \|(X_i,Y_i) \mid \mathcal{E} - (X_i,Y_i)\|_{\mathrm{TV}}^2 \;\lesssim\; \sum_{i=1}^n \mathrm{KL}\big(\mu_{X_i,Y_i|\mathcal{E}} \,\|\, \mu_{X_i,Y_i}\big).$$

Using the chain rule for KL divergence on product distributions (or, equivalently, the non-negativity of mutual information between coordinates), one can show

$$\sum_{i=1}^n \mathrm{KL}\big(\mu_{X_i,Y_i|\mathcal{E}} \,\|\, \mu_{X_i,Y_i}\big) \;\le\; \mathrm{KL}\big(\mu_{X^n,Y^n|\mathcal{E}} \,\|\, \mu_{X^n,Y^n}\big),$$

where (X^n, Y^n) denotes the entire transcript. Finally,

$$\mathrm{KL}\big(\mu_{X^n,Y^n|\mathcal{E}} \,\|\, \mu_{X^n,Y^n}\big) = \mathbb{E}_{x^n,y^n\sim\mu(\cdot|\mathcal{E})}\left[\log\frac{\mu(x^n,y^n \mid \mathcal{E})}{\mu(x^n,y^n)}\right]$$
$$= \mathbb{E}_{x^n,y^n\sim\mu(\cdot|\mathcal{E})}\left[\log\frac{\mu(x^n,y^n,\mathcal{E})}{\mu(x^n,y^n)\mathbb{P}[\mathcal{E}]}\right] \le \log\frac{1}{\mathbb{P}[\mathcal{E}]},$$

since $\mu(\mathcal{E} \mid x^n, y^n) \le 1$. ∎

11.6 Correlated sampling

The following "correlated sampling" fact is used to let Alice and Bob sample the same random variable from two nearby conditional distributions, using only public randomness.

FACT 11.12 (Correlated sampling (Holenstein)). *Let Z be a random variable and let X, Y be two pieces of side information. If*

$$\|Z \mid X - Z \mid Y\|_{\mathrm{TV}} \le \varepsilon,$$

then there is a public-coin protocol such that:

- *Alice outputs $Z_A \sim (Z \mid X)$ and Bob outputs $Z_B \sim (Z \mid Y)$ (exactly); and*
- *$\mathbb{P}[Z_A = Z_B] \ge 1 - 2\varepsilon$.*

Proof sketch (rejection sampling). Interpret the public random string as an infinite sequence of i.i.d. pairs

$$(z_1, r_1), (z_2, r_2), \ldots$$

where each z_i is uniform over the support of Z and each r_i is uniform in $[0, 1]$. Alice scans the list and takes the first z_i such that $r_i \le \mathbb{P}[Z = z_i \mid X]$; Bob does the analogous rule with $\mathbb{P}[Z = z_i \mid Y]$.

A short calculation shows when scanning index i, for Ω the probability space of Z,

$$\mathbb{P}[\text{Alice or Bob accepts } (z_i, r_i)] = \sum_z \frac{1}{|\Omega|}\max\{\mathbb{P}[Z = z \mid X], \mathbb{P}[Z = z \mid Y]\},$$

while

$$\mathbb{P}[\text{Alice and Bob both accept } (z_i, r_i)] = \sum_z \frac{1}{|\Omega|}\min\{\mathbb{P}[Z = z \mid X], \mathbb{P}[Z = z \mid Y]\}.$$

Therefore,

$$\mathbb{P}[Z_A = Z_B] \ \ge\ \frac{\sum_z \min\{\mathbb{P}[Z = z \mid X], \mathbb{P}[Z = z \mid Y]\}}{\sum_z \max\{\mathbb{P}[Z = z \mid X], \mathbb{P}[Z = z \mid Y]\}} = \frac{1 - \|Z \mid X - Z \mid Y\|_{\mathrm{TV}}}{1 + \|Z \mid X - Z \mid Y\|_{\mathrm{TV}}} \ \ge\ 1 - 2\varepsilon.$$

∎

11.7 A key lemma behind the embedding

We now state (and prove) the key quantitative lemma that underlies the embedding step.

Setup. Let G be a 2P1R game and consider the repeated game $G^{\otimes n}$ under a fixed deterministic strategy. We view the questions as random variables $(X_1, Y_1), \ldots, (X_n, Y_n) \sim \mu^{\otimes n}$ and the answers as random variables $(A_1, B_1), \ldots, (A_n, B_n)$ produced by the strategy.

Fix a parameter $k \leq n$ (think: $k = \varepsilon n$). We will condition on winning the last $n - k$ coordinates:

$$E := W_{k+1} \wedge \cdots \wedge W_n.$$

(Other choices of E are possible; this is the one used in the notes.)

Define auxiliary "dependency-breaking" randomness as follows. For each $i \in [k]$, sample an independent fair coin $D_i \in \{0, 1\}$ and define

$$(D_i = 1): \quad U_i := X_i,\ \bar{U}_i := Y_i, \qquad\qquad (D_i = 0): \quad U_i := Y_i,\ \bar{U}_i := X_i.$$

The variable $\bar{U}_i$ is the question we reveal publicly, while U_i is the "missing" question to be sampled later by one of the players.

Finally define

$$\begin{aligned} T &:= (X_{k+1}, \ldots, X_n,\ Y_{k+1}, \ldots, Y_n,\ D_1, \ldots, D_k,\ \bar{U}_1, \ldots, \bar{U}_k), \\ V &:= (A_{k+1}, \ldots, A_n,\ B_{k+1}, \ldots, B_n). \end{aligned}$$

Key lemma.

Lemma 11.13. *With the notation above,*

$$\sum_{j=1}^{k} \|(T, U_j, V) \mid \mathcal{E} - (T, V) \mid \mathcal{E} \ \cdot\ (U_j \mid T)\|_{\mathrm{TV}} \ \leq\ \sqrt{k}\,\sqrt{(n-k)\,\log(|\Sigma_A|\,|\Sigma_B|) + \log \frac{1}{\mathbb{P}[\mathcal{E}]}}.$$

Remark 11.14. The left-hand side measures, on average over j, how far U_j is from being conditionally independent of V given T under the conditioning event E. If this distance is small, then (for a typical j) Alice and Bob can use Fact 11.12 to *agree* on a sample of the missing randomness needed to embed coordinate j.

How this enables correlated sampling (as in the notes). For each $j \in [k]$, define

$$\Delta_j \ :=\ \|(T, U_j, V) \mid \mathcal{E} - (T, V) \mid \mathcal{E} \cdot (U_j \mid T)\|_{\mathrm{TV}}\,.$$

Lemma 11.13 shows $\sum_{j=1}^{k} \Delta_j$ is small (for suitable parameters), so for a typical coordinate j we have $\Delta_j \ll 1$.

Fix such a coordinate j and consider, for instance, the case $D_j = 0$ (so $\bar{U}_j = X_j$ is publicly revealed and $U_j = Y_j$ is the "missing" question). Write T^{-j} for the variable T with the components $(D_j, \bar{U}_j)$ removed. A straightforward manipulation of conditional distributions (the algebra carried out on the notes) shows that small Δ_j implies that the joint distribution of (T^{-j}, V, X_j, Y_j) conditioned on $\mathcal{E}$ is close to the product form in which (X_j, Y_j) has its original marginal:

$$\left\|\mu_{T^{-j}, V, X_j, Y_j \mid \mathcal{E},\, D_j = 0} - \mu_{T^{-j}, V \mid Y_j, \mathcal{E},\, D_j = 0} \cdot \mu_{X_j, Y_j}\right\|_{\mathrm{TV}} \ \leq\ 2\Delta_j.$$

From this one can further derive that

$$\left\|\mu_{T^{-j},V|X_j,\mathcal{E}} \cdot \mu_{X_j,Y_j} - \mu_{T^{-j},V|Y_j,\mathcal{E}} \cdot \mu_{X_j,Y_j}\right\|_{\mathrm{TV}} \;\le\; O(\Delta_j).$$

Summing over j and averaging, there exists at least one coordinate j for which $\mu_{T^{-j},V|X_j,\mathcal{E}}$ and $\mu_{T^{-j},V|Y_j,\mathcal{E}}$ are close in total variation.

For such a coordinate j, Alice (who knows X_j) and Bob (who knows Y_j) can apply the correlated sampling protocol from Fact 11.12 to *agree* on a common value of (T^{-j}, V) using only public randomness. Conditioned on this common value, the remaining (unrevealed) questions become independent in the right way, so Alice and Bob can sample the missing questions locally and then run the repeated-game strategy to obtain an answer for coordinate j. This implements the desired embedding step.

Proof. Fix $j \in [k]$. Expanding the definition of total variation and conditioning,

$$\|(T, U_j, V) \mid \mathcal{E} - (T, V) \mid \mathcal{E} \cdot (U_j \mid T)\|_{\mathrm{TV}} = \mathbb{E}_{(t,v)\sim(T,V)|\mathcal{E}}\left[\|U_j \mid (t, v, \mathcal{E}) - U_j \mid t\|_{\mathrm{TV}}\right].$$

Summing over j and applying Cauchy–Schwarz,

$$\begin{aligned}\sum_{j=1}^{k} \|(T, U_j, V) \mid \mathcal{E} - (T, V) \mid \mathcal{E} \cdot (U_j \mid T)\|_{\mathrm{TV}} &\le \mathbb{E}_{(t,v)\sim(T,V)|\mathcal{E}}\left[\sum_{j=1}^{k} \|U_j \mid (t, v, \mathcal{E}) - U_j \mid t\|_{\mathrm{TV}}\right] \\ &\le \mathbb{E}_{(t,v)\sim(T,V)|\mathcal{E}}\left[\sqrt{k}\sqrt{\sum_{j=1}^{k} \|U_j \mid (t, v, \mathcal{E}) - U_j \mid t\|_{\mathrm{TV}}^2}\right].\end{aligned}$$

By Pinsker (Fact 11.7), $\|\cdot - \cdot\|_{\mathrm{TV}}^2 \lesssim \mathrm{KL}(\cdot\|\cdot)$, so

$$\sum_{j=1}^{k} \|U_j \mid (t, v, \mathcal{E}) - U_j \mid t\|_{\mathrm{TV}}^2 \;\lesssim\; \sum_{j=1}^{k} \mathrm{KL}\big(\mu_{U_j|t,v,\mathcal{E}} \,\|\, \mu_{U_j|t}\big).$$

Using the chain rule for KL divergence (or convexity of KL), the sum over coordinates is bounded by the KL divergence of the joint variable $U := (U_1, \ldots, U_k)$:

$$\sum_{j=1}^{k} \mathrm{KL}\big(\mu_{U_j|t,v,\mathcal{E}} \,\|\, \mu_{U_j|t}\big) \;\le\; \mathrm{KL}\big(\mu_{U|t,v,\mathcal{E}} \,\|\, \mu_{U|t}\big).$$

Plugging this in and applying Jensen's inequality to the concave function $\sqrt{\cdot}$ yields

$$\sum_{j=1}^{k} \|(T, U_j, V) \mid \mathcal{E} - (T, V) \mid \mathcal{E} \cdot (U_j \mid T)\|_{\mathrm{TV}} \;\lesssim\; \sqrt{k}\sqrt{\mathbb{E}_{(t,v)\sim(T,V)|\mathcal{E}}\Big[\mathrm{KL}\big(\mu_{U|t,v,\mathcal{E}} \,\|\, \mu_{U|t}\big)\Big]}.$$

It remains to bound the expectation of the KL term. We decompose it into "correlation" (mutual information) and "conditioning":

$$\begin{aligned}\mathbb{E}_{t,v|\mathcal{E}}\Big[\mathrm{KL}\big(\mu_{U|t,v,\mathcal{E}} \,\|\, \mu_{U|t}\big)\Big] &\le \mathbb{E}_{t,v|\mathcal{E}}\Big[\mathrm{KL}\big(\mu_{U|t,v,\mathcal{E}} \,\|\, \mu_{U|t,\mathcal{E}}\big)\Big] + \mathbb{E}_{t|\mathcal{E}}\Big[\mathrm{KL}\big(\mu_{U|t,\mathcal{E}} \,\|\, \mu_{U|t}\big)\Big] \\ &= \mathrm{I}(U; V \mid T, \mathcal{E}) \;+\; \mathbb{E}_{t|\mathcal{E}}\Big[\mathrm{KL}\big(\mu_{U|t,\mathcal{E}} \,\|\, \mu_{U|t}\big)\Big].\end{aligned}$$

The mutual information term is bounded by the entropy of V (Fact 11.10):

$$\mathrm{I}(U; V \mid T, \mathcal{E}) \le \mathrm{H}(V \mid T, \mathcal{E}) \le \mathrm{H}(V \mid \mathcal{E}) \le (n-k)\, \log(|\Sigma_A|\, |\Sigma_B|),$$

since V is a string of length $n-k$ over $\Sigma_A \times \Sigma_B$.

For the second term, conditioning on E can reveal at most $\log(1/\mathbb{P}[\mathcal{E}])$ bits. Formally, for every fixed t,

$$\mathrm{KL}\big(\mu_{U|t,\mathcal{E}} \,\|\, \mu_{U|t}\big) = \mathbb{E}_{u \sim \mu(\cdot|t,\mathcal{E})}\left[\log \frac{\mu(u \mid t, \mathcal{E})}{\mu(u \mid t)}\right] \le \log \frac{1}{\mathbb{P}[\mathcal{E} \mid T = t]},$$

and therefore

$$\mathbb{E}_{t|\mathcal{E}}\Big[\mathrm{KL}\big(\mu_{U|t,\mathcal{E}} \,\|\, \mu_{U|t}\big)\Big] \le \mathbb{E}_{t|\mathcal{E}}\left[\log \frac{1}{\mathbb{P}[\mathcal{E} \mid T = t]}\right] = \log \frac{1}{\mathbb{P}[\mathcal{E}]}.$$

Combining the bounds proves the lemma (up to the universal constant hidden in Pinsker's inequality). ∎

Lemma 11.13, together with the correlated sampling fact (Fact 11.12), is the heart of the embedding argument and leads to the exponential decay in Theorem 11.3.