Appendix to "New Sample Complexity Bounds for Learning a Phylogeny from Multiple Loci"

Gautam Dasarathy
Electrical Engineering
University of Wisconsin - Madison
Madison, Wisconsin 53706
dasarathy@wisc.edu

Robert Nowak
Electrical Engineering
University of Wisconsin - Madison
Madison, Wisconsin 53706
nowak@ece.wisc.edu

Sebastien Roch
Mathematics
University of Wisconsin - Madison
Madison, Wisconsin 53706
roch@math.wisc.edu

This document will serve as a companion to the submission titled *New Sample Complexity Bounds for Learning Phylogeny from Multiple Loci* and all theorem numbers refer to the theorem numbers in that document.

I. Proof of Theorem 1

Recall that for any pair of leaves $A, B \in L$, we define

$$\widehat{p}_{AB} = \frac{1}{mk} \sum_{i\in[m],j\in[k]} \mathbb{1}\{\chi_A^{ij} \neq \chi_B^{ij}\}. \tag{1}$$

Theorem 1. *$\{\mathbb{E}[\widehat{p}_{AB}]\}_{A,B\in L}$[†] forms an ultrametric with respect to the true species tree S. In fact, for any triple $A, B, C \in L$ with the topology $((A, B), C)$ in S, we have*

$$\mathbb{E}[\widehat{p}_{AC}] = \mathbb{E}[\widehat{p}_{BC}] > \mathbb{E}[\widehat{p}_{AB}] + \frac{3e^{-\frac{4}{3}\mu\tau_{AC}}\mu f}{8\mu+3}. \tag{2}$$

Proof: Suppose that $A, B, C \in L$ are three arbitrary leaves of the species tree with the topology $((A, B), C)$. By definition, we have that

$$\mathbb{E}[\widehat{p}_{AC}] = \mathbb{E}\left[\frac{3}{4}\left(1 - e^{-\frac{4}{3}\delta_{AC}}\right)\right],$$

where δ_{AC} is the random gene tree distance that satisfies $\delta_{AC} = \mu\tau_{AC} + 2\mu Z$ with $Z \sim \text{Exp}(1)$. Therefore, we have

$$\begin{aligned}
\mathbb{E}[\widehat{p}_{AC}] - \mathbb{E}[\widehat{p}_{AB}] &= -\frac{3}{4}e^{-\frac{4}{3}\mu\tau_{AC}}\mathbb{E}\left[e^{-\frac{8}{3}\mu Z}\right] + -\frac{3}{4}e^{-\frac{4}{3}\mu\tau_{AB}}\mathbb{E}\left[e^{-\frac{8}{3}\mu Z}\right] \\
&\overset{(a)}{=} \frac{3\left(e^{-\frac{4}{3}\mu\tau_{AB}} - e^{-\frac{4}{3}\mu\tau_{AC}}\right)}{4(\frac{8}{3}\mu+1)} \\
&\overset{(b)}{\geq} \frac{3e^{-\frac{4}{3}\mu\tau_{AC}}\mu f}{(8\mu+3)},
\end{aligned}$$

where (a) follows from the fact that if $X \sim$Exp(1), for any $\alpha > 0$, $\mathbb{E}[e^{-\alpha X}] = (\alpha+1)^{-1}$ and (b) follows from the fact that for any $\alpha > 0$ and $x < y$, $\frac{e^{-\alpha x}}{\alpha} - \frac{e^{-\alpha y}}{\alpha} \geq (y-x)\, e^{-\alpha y}$. Proceeding similarly, It can be seen that $\mathbb{E}[\widehat{p}_{AC}] = \mathbb{E}[\widehat{p}_{BC}]$. This concludes the proof. ∎

II. Proof of Theorem 2

We next prove Theorem 2 which guarantees that S can be reliably recovered by using a standard distance-based algorithm like UPGMA or bottom-up agglomerative clustering with $\{\widehat{p}_{AB}\}_{A,B\in L}$ as a dissimilarity measure for L.

Theorem 2. *Given an $\epsilon > 0$, using UPGMA on L with the dissimilarity measure $\{\widehat{p}_{AB}\}_{A,B\in L}$ results in the correct tree S being output with probability no less than $1-\epsilon$ as long as the number of gene trees m, and the number of samples per gene tree k satisfy*

$$m \geq \frac{16\, e^{\frac{8}{3}\mu\Delta}(8\mu+3)^2}{9\mu^2 f^2} \log\left(\frac{4\binom{n}{3}}{\epsilon}\right) \quad \textit{and} \quad k \geq 1 \tag{3}$$

[†]Unless otherwise noted, expectations will be with respect all the randomness present.

Proof: Let us consider the probability of error.

$$\begin{aligned}\mathbb{P}\left[\text{Error}\right] &= \mathbb{P}\left[\bigcup_{((A,B),C)\in\binom{L}{3}} \left\{\text{The triple } ((A,B),C) \text{ is such that } \widehat{p}_{AB} > \widehat{p}_{AC}\right\}\right] \\ &\leq \sum_{((A,B),C)\in\binom{L}{3}} \mathbb{P}\left[\widehat{p}_{AB} > \widehat{p}_{AC}\right] \end{aligned} \tag{4}$$

Now, if we define $\alpha_{\text{um}} = \frac{3e^{-\frac{4}{3}\Delta}\mu f}{(8\mu+3)}$, then Theorem 1 guarantees that for an arbitrary triple $((A,B),C)$,

$$\begin{aligned}\mathbb{P}\left[\widehat{p}_{AB} - \widehat{p}_{AC} > 0\right] &\leq \mathbb{P}\left[\widehat{p}_{AB} - \mathbb{E}\left[p_{AB}\right] - \widehat{p}_{AC} + \mathbb{E}\left[p_{AC}\right] > \alpha_{\text{um}}\right] \\ &\leq \mathbb{P}\left[\widehat{p}_{AB} - \mathbb{E}\left[p_{AB}\right] > \frac{\alpha_{\text{um}}}{2}\right] + \mathbb{P}\left[\mathbb{E}\left[p_{AC}\right] - \widehat{p}_{AC} > \frac{\alpha_{\text{um}}}{2}\right] \end{aligned} \tag{5}$$

Let us first look at the first term in (5). The second one will follow similarly.

$$\begin{aligned}&\mathbb{P}\left[\widehat{p}_{AB} - \mathbb{E}[p_{AB}] > \alpha_{\text{um}}/2\right] \\ &\overset{(a)}{=} \mathbb{E}\left[\mathbb{P}\left(\widehat{p}_{AB} - \mathbb{E}\left[p_{AB}\right] > \frac{\alpha_{\text{um}}}{2}\middle|\{\delta_{AB}^{(i)}\}_{i\in[m]}\right)\right] \\ &\leq \mathbb{E}\left[\mathbb{P}\left(\widehat{p}_{AB} - \frac{1}{m}\sum_{i\in[m]} p_{AB}^{(i)} > \frac{\alpha_{\text{um}}}{4}\middle|\{\delta_{AB}^{(i)}\}_{i\in[m]}\right) + \mathbb{P}\left(\frac{1}{m}\sum_{i\in[m]} p_{AB}^{(i)} - \mathbb{E}\left[p_{AB}\right] > \frac{\alpha_{\text{um}}}{4}\middle|\{\delta_{AB}^{(i)}\}_{i\in[m]}\right)\right] \\ &= \mathbb{E}\left[\mathbb{P}\left(\widehat{p}_{AB} - \frac{1}{m}\sum_{i\in[m]} p_{AB}^{(i)} > \frac{\alpha_{\text{um}}}{4}\middle|\{\delta_{AB}^{(i)}\}_{i\in[m]}\right)\right] + \mathbb{P}\left(\frac{1}{m}\sum_{i\in[m]} p_{AB}^{(i)} - \mathbb{E}\left[p_{AB}\right] > \frac{\alpha_{\text{um}}}{4}\right), \end{aligned}$$

where in (a), $\delta_{AB}^{(i)}$ is the random distance between leaves A and B in gene tree $\mathcal{G}^{(i)}$. Each term in the last equation can now be upper bounded using Hoeffding's inequality.

$$\mathbb{E}\left[\mathbb{P}\left[\frac{1}{mk}\sum_{i=1}^{m}\sum_{j=1}^{k} X_{AB}^{ij} - \frac{1}{m}\sum_{i=1}^{m} p_{AB}^{(i)} > \frac{\alpha_{\text{um}}}{4}\middle|\left\{d_{AB}^{(i)}\right\}\right]\right] \leq e^{-mk\alpha_{\text{um}}^2/16}$$

$$\mathbb{P}\left(\frac{1}{m}\sum_{i\in[m]} p_{AB}^{(i)} - \mathbb{E}\left[p_{AB}\right] > \frac{\alpha_{\text{um}}}{4}\right) \leq e^{-m\alpha_{\text{um}}^2/16}$$

Substituting these in (4), we have

$$\begin{aligned}\mathbb{P}\left[\text{Error}\right] &\leq \sum_{((AB)C)\in\binom{L}{3}} \mathbb{P}\left[\widehat{p}_{AB} > \widehat{p}_{AC}\right] \\ &\leq \sum_{((AB)C)\in\binom{L}{3}} 2\left(e^{-mk\alpha_{\text{um}}^2/16} + e^{-m\alpha_{\text{um}}^2/16}\right) \\ &\leq \binom{n}{3} 2\left(e^{-mk\alpha_{\text{um}}^2/16} + e^{-m\alpha_{\text{um}}^2/16}\right) \end{aligned}$$

Therefore, the probability of error can be made less than ϵ if we pick m and k as shown in (3). ■

III. Proof of Theorem 3

Recall that we define $d_{AB} = -\frac{3}{4}\log\left(1 - \frac{4}{3}\mathbb{E}\left[\widehat{p}_{AB}\right]\right)$ and Theorem 3 tells us that these distances form an additive metric with respect to S. [1].

Theorem 3. *The set of dissimilarities $\{d_{AB}\}_{A,B\in L}$ forms an additive metric with respect to S. In fact, suppose the leaves $A, B, C, D \in L$ are such that either $((A,B),(C,D))$ or $(((A,B),C),D)$ holds with respect to S, then*

$$d_{AC} + d_{BD} = d_{AD} + d_{BC} > d_{AB} + d_{CD} + \alpha_{\text{add}},$$

where $\alpha_{\text{add}} = \log\left(\frac{8}{3}\mu_*(1 - e^{-f}) + 1\right) > 0$.

[1]Recall that this means that the four point condition holds, i.e., for a quadruple of leaves A, B, C, D that are such that $((A,B),(C,D))$ or $(((A,B),C),D)$, the distances satisfy $d_{AB} + d_{CD} \leq d_{AC} + d_{BD} = d_{AD} + d_{BC}$

Fig. 1: Pictures showing the random variables and internal nodes used in Proof of Theorem 3

Proof: As a first step we note that for any 4 leaves of the species tree A, B, C, D, there are only 2 possible topologies with respect to S (upto relabeling): (a) $((A,B),(C,D))$ and (b) $(((A,B),C),D)$. We will consider each case separately in what follows.

Case (a): In order to tackle the first case, we will use the notation from Figure 1a, which shows the species tree S restricted to the leaves A, B, C, D. Let o_1, o_2 and o_3 be the common ancestors of (A,B), (C,D) and (A,C) respectively. Let $\mathcal{E}_{AB}$ be the event that the lineages corresponding to A and B coalesce in the segment (o_1, o_3) of the tree in Figure 1a and let $\overline{\mathcal{E}_{AB}}$ be the event that this does not occur. Similarly, we define the events $\mathcal{E}_{CD}$ and $\overline{\mathcal{E}_{CD}}$. To reduce notational clutter, for $w, v \in S$, we will write μ_{wv} to denote $\sum_{e \in \pi^S_{wv}} \mu_e \tau_e$. Now, for leaves $X, Y \in L$, let Z_{XY} denote the random quantity $\frac{1}{2}(\delta_{XY} - \mu_{XY})$, i.e., it is the effective (mutation rate adjusted) coalescent time after the lineages corresponding to X and Y find themselves in a common population. It is easy to check that the quantities $Z_{AB} - \mu_{o_1 o_3} \mid \overline{\mathcal{E}_{AB}}$, $Z_{CD} - \mu_{o_2 o_3} \mid \overline{\mathcal{E}_{AB}}$, Z_{AC}, and Z_{CD} have the same distribution. Let Z denote this common random variable; this is shown diagrammatically in Figure 1a. Now, we will use the fact that by definition, $\delta_{AB} = \mu_{AB} + 2Z_{AB}$ and $\delta_{CD} = \mu_{CD} + 2Z_{CD}$, and the fact that conditioned on $\mathcal{E}_{AB}$ (resp. $\mathcal{E}_{CD}$), $Z_{AB} \leq \mu_{o_1 o_3}$ (resp. $Z_{CD} \leq \mu_{o_2 o_3}$) to obtain the lower bounds

$$\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AB}}\right] \geq e^{-\frac{4}{3}\mu_{AB}} \left\{ \mathbb{E}\left[e^{-\frac{8}{3}\mu_{o_1 o_3}} \middle| \mathcal{E}_{AB}\right] \mathbb{P}\left(\mathcal{E}_{AB}\right) + e^{-\frac{8}{3}\mu_{o_1 o_3}} \mathbb{E}\left[e^{-\frac{8}{3}Z}\right] \mathbb{P}\left(\overline{\mathcal{E}_{AB}}\right) \right\} \tag{6}$$

$$\mathbb{E}\left[e^{-\frac{4}{3}\delta_{CD}}\right] \geq e^{-\frac{4}{3}\mu_{CD}} \left\{ \mathbb{E}\left[e^{-\frac{8}{3}\mu_{o_2 o_3}} \middle| \mathcal{E}_{CD}\right] \mathbb{P}\left(\mathcal{E}_{CD}\right) + e^{-\frac{8}{3}\mu_{o_2 o_3}} \mathbb{E}\left[e^{-\frac{8}{3}Z}\right] \mathbb{P}\left(\overline{\mathcal{E}_{CD}}\right) \right\} \tag{7}$$

On the other hand, notice that $\delta_{AC} = \mu_{AC} + 2Z$ and $\delta_{BD} = \mu_{BD} + 2Z$. Therefore, we have

$$\begin{aligned}
\frac{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AB}}\right]\mathbb{E}\left[e^{-\frac{4}{3}\delta_{CD}}\right]}{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AC}}\right]\mathbb{E}\left[e^{-\frac{4}{3}\delta_{BD}}\right]} &\geq \frac{e^{-\frac{4}{3}\mu_{AB}} \left\{ \mathbb{E}\left[e^{-\frac{8}{3}\mu_{o_1 o_3}} \middle| \mathcal{E}_{AB}\right] \mathbb{P}\left(\mathcal{E}_{AB}\right) + e^{-\frac{8}{3}\mu_{o_1 o_3}} \mathbb{E}\left[e^{-\frac{8}{3}Z}\right] \mathbb{P}\left(\overline{\mathcal{E}_{AB}}\right) \right\}}{e^{-\frac{4}{3}\mu_{AC}} \mathbb{E}\left[e^{-\frac{8}{3}Z}\right]} \\
&\quad \times \frac{e^{-\frac{4}{3}\mu_{CD}} \left\{ \mathbb{E}\left[e^{-\frac{8}{3}\mu_{o_2 o_3}} \middle| \mathcal{E}_{CD}\right] \mathbb{P}\left(\mathcal{E}_{CD}\right) + e^{-\frac{8}{3}\mu_{o_2 o_3}} \mathbb{E}\left[e^{-\frac{8}{3}Z}\right] \mathbb{P}\left(\overline{\mathcal{E}_{C}}\right) \right\}}{e^{-\frac{4}{3}\mu_{BD}} \mathbb{E}\left[e^{-\frac{8}{3}Z}\right]} \\
&\geq \left[\frac{\mathbb{P}\left(\mathcal{E}_{AB}\right)}{\mathbb{E}\left[e^{-\frac{8}{3}Z}\right]} + \mathbb{P}\left(\overline{\mathcal{E}_{AB}}\right) \right] \times \left[\frac{\mathbb{P}\left(\mathcal{E}_{CD}\right)}{\mathbb{E}\left[e^{-\frac{8}{3}Z}\right]} + \mathbb{P}\left(\overline{\mathcal{E}_{CD}}\right) \right] \\
&\geq \left[\frac{8}{3}\mu_* \left(1 - e^{-f}\right) + 1 \right]^2,
\end{aligned} \tag{8}$$

where in the last step we have used the fact that the random variable Z stochastically dominates the random variable $\mu_* \tilde{Z}$, where $\tilde{Z} \sim \text{Exp}(1)$ and that that $\mathbb{P}\left[\mathcal{E}_{XY}\right] \geq 1 - e^{-f}$ for each pair of leaves $X, Y \in L$. Next, we consider Case (b).

Case (b) : Here, we will write o_1, o_2, o_3 to denote the most recent common ancestors of (A,B), (A,C) and (A,D) respectively. Again we will use notation from the previous case for random variables of the form $Z_{XY}, X, Y \in L$. In this case, we let $\mathcal{E}_{AB}$

denote the event that the lineages corresponding to A and B coalesce in the branch (o_1, o_2) in Figure 1b. Again, from the properties of the multispecies coalescent, we known that $Z_{AB} - \mu_{o_1 o_2} \mid \overline{\mathcal{E}_{AB}}$ and Z_{AC} have the same distribution and we let Z_1 denote a random variable with this common distribution. Similarly Z_{CD} and Z_{BD} have the same distribution and we let Z_2 denote a random variable with this distribution. Performing a similar calculation as before leads us to

$$\begin{aligned}\frac{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AB}}\right]\mathbb{E}\left[e^{-\frac{4}{3}\delta_{CD}}\right]}{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AC}}\right]\mathbb{E}\left[e^{-\frac{4}{3}\delta_{BD}}\right]} &= \frac{\mathbb{E}\left[e^{\frac{8}{3}(\mu_{o_1o_2}-Z_{AB})}\middle|\mathcal{E}_{AB}\right]}{\mathbb{E}\left[e^{-\frac{8}{3}Z_1}\right]}\mathbb{P}\left[\mathcal{E}_{AB}\right] + \mathbb{P}\left[\overline{\mathcal{E}_{AB}}\right] \\ &\overset{(a)}{\geq} \frac{\mathbb{P}\left[\mathcal{E}_{AB}\right]}{\mathbb{E}\left[e^{-\frac{8}{3}Z_1}\right]} + \mathbb{P}\left[\overline{\mathcal{E}_{AB}}\right] \\ &\overset{(b)}{\geq} \left[\frac{8}{3}\mu_*(1-e^{-f})+1\right], \end{aligned} \tag{9}$$

where in (a), as in the previous case, we use the fact that conditioned on $\mathcal{E}_{AB}$, $\mu_{o_1o_2} \geq Z_{AB}$ and in (b) $\mathbb{E}\left[e^{-\frac{8}{3}Z_1}\right] \leq \frac{1}{\frac{8}{3}\mu_*+1}$ and that $\mathbb{P}\left[\mathcal{E}_{XY}\right] \geq 1-e^{-f}$ for all pairs of leaves $X, Y \in L$. Taking logarithms on either side of (8) and (9) gives us the result that $d_{AC}+d_{BD} \geq d_{AB}+d_{CD}+\log\left(\frac{8}{3}\mu_*\left(1-e^{-f}\right)+1\right)$. Using a similar procedure, one can show that $d_{AC}+d_{BD} = d_{AD}+d_{BC}$. This proves that $\{d_{AB}\}_{A,B\in L}$ is an additive metric with respect to S. ∎

IV. Proof of Theorem 4

We will now prove the last main result in our paper that shows that the above result can be used to obtain sample complexity bounds when one has access to molecular data. In particular, we propose the following measure of dissimilarity from the samples (with $\widehat{p}_{AB}$ as defined above)

$$\widehat{d}_{AB} \triangleq -\frac{3}{4}\log\left(1-\frac{4}{3}\widehat{p}_{AB}\right). \tag{10}$$

In light of Theorem 3, we propose the following tree reconstruction procedure in the main paper: use any distance algorithm (like Neighbor Joining [**?**]) which returns an additive tree with $\{\widehat{d}_{AB}\}_{A,B\in L}$ defined as in (10) as input. We then have the following result.

Theorem 4. *For any $\epsilon > 0$, the above method succeeds in reconstructing (the unrooted version of) S with probability at least $1-\epsilon$ as long as m and k satisfy*

$$m \geq \frac{e^{\frac{8\mu^*\Delta}{3}}(\frac{8}{3}\mu^*+1)^2(8+4\alpha_{\text{add}})^2}{\alpha_{\text{add}}^2}\log\left(\frac{4\binom{n}{4}}{\epsilon}\right) \quad \text{and} \quad k \geq 1, \tag{11}$$

where $\alpha_{\text{add}} = \log\left(\frac{8}{3}\mu_*(1-e^{-f})+1\right)$. In the limit as $f \to 0$, the quantity on the right side approaches $\frac{e^{\frac{8\mu^*\Delta}{3}}(8\mu^*+3)^2}{\mu_*^2 f^2}\log\left(\frac{4\binom{n}{3}}{\epsilon}\right)$.

Proof: Notice that it suffices to show that for any four leaves A, B, C, D such that $\tau_{AB}+\tau_{CD} \leq \tau_{AC}+\tau_{BD} = \tau_{AD}+\tau_{BC}$ (i.e., $((A,B),(C,D))$ or $(((A,B),C),D)$ holds with respect to S), the 4-point condition is satisfied by $\hat{d}$ with high probability.

We begin by setting $\alpha_{\text{add}} \triangleq \log\left(4\mu_*(1-e^{-f})+1\right)$ and observing that Theorem 3 tells us that

$$\begin{aligned}\mathbb{P}\left[\widehat{d}_{AB}+\widehat{d}_{CD}-\widehat{d}_{AC}-\widehat{d}_{BD}>0\right] \leq\ & \mathbb{P}\left[\widehat{d}_{AB}-d_{AB}>\frac{\alpha_{\text{add}}}{4}\right]+\mathbb{P}\left[\widehat{d}_{CD}-d_{CD}>\frac{\alpha_{\text{add}}}{4}\right] \\ &+\mathbb{P}\left[d_{AC}-\widehat{d}_{AC}>\frac{\alpha_{\text{add}}}{4}\right]+\mathbb{P}\left[d_{BD}-\widehat{d}_{BD}>\frac{\alpha_{\text{add}}}{4}\right].\end{aligned}$$

We will bound one of the 4 terms above and the bounds for the remaining terms will follow similarly. Towards this end, let $p_{AB}^{(i)}$ denote the random quantity $\frac{3}{4}\left(1-e^{-\frac{4}{3}\delta_{AB}^{(i)}}\right)$ and let $\mathcal{E}(\delta)$ denote the event that $\left|\frac{1}{m}\sum_{i\in[m]}p_{AB}^{(i)}-\mathbb{E}p_{AB}\right|>\delta$. Also, to

reduce notational clutter, we let $\ell(x)$ denote the function $-\frac{3}{4}\log(1-\frac{4}{3}x)$. Now, observe that

$$\begin{aligned}
\mathbb{P}\left[\widehat{d}_{AB} - d_{AB} > \frac{\alpha_{\text{add}}}{4}\right] &= \mathbb{P}\left[\ell(\widehat{p}_{AB}) - \ell(\mathbb{E}p_{AB}) > \frac{\alpha_{\text{add}}}{4}\right] \\
&\leq \underbrace{\mathbb{P}\left[\mathcal{E}(\delta)\right]}_{(a)} + \underbrace{\mathbb{E}\left[\mathbb{P}\left[\ell\left(\widehat{p}_{AB}\right) - \ell\left(\frac{1}{m}\sum_{i\in[m]} p_{AB}^{(i)}\right) > \frac{\alpha_{\text{add}}}{8}\middle|\left\{\delta_{AB}^{(i)}\right\}_{i\in[m]}, \mathcal{E}(\delta)^c\right]\right]}_{(b)} \\
&\quad + \underbrace{\mathbb{E}\left[\mathbb{P}\left[\ell\left(\frac{1}{m}\sum_{i\in[m]} p_{AB}^{(i)}\right) - \ell\left(\mathbb{E}\left[p_{AB}\right]\right) > \frac{\alpha_{\text{add}}}{8}\middle|\left\{\delta_{AB}^{(i)}\right\}_{i\in[m]}, \mathcal{E}(\delta)^c\right]\right]}_{(c)}
\end{aligned}$$

We will now pick $\delta > 0$ so that (c) equals 0. To do this, we observe that from the properties of the multispecies coalescent we have

$$\begin{aligned}
\mathbb{E}\left[p_{AB}\right] &= \mathbb{E}\left[\frac{3}{4}(1 - e^{-\frac{4}{3}\delta_{AB}})\right] \\
&= \frac{3}{4}\left(1 - e^{-\frac{4}{3}\mu_{AB}}\mathbb{E}\left[e^{-\frac{8}{3}Z_{AB}}\right]\right) \\
&\leq \frac{3}{4}\left(1 - \frac{e^{-\frac{4}{3}\mu^*\Delta}}{\frac{8}{3}\mu^* + 1}\right)
\end{aligned}$$

Therefore, if one picks $\delta = \frac{\alpha_{\text{add}} e^{-\frac{4}{3}\mu^*\Delta}}{(8+4\alpha_{\text{add}})(\frac{8}{3}\mu^*+1)}$, then it can be seen that (c) is zero. Now, we consider (b). Let $\mathcal{E}_2(\delta_2)$ denote the event that $\left|\widehat{p}_{AB} - \frac{1}{m}\sum_{i\in[m]} p_{AB}^{(i)}\right| > \delta_2$ given $\left\{d_{AB}^{(i)}\right\}_{i\in[m]}$. We can proceed as above and set $\delta_2 = \delta$, which gives us the following

$$\begin{aligned}
\mathbb{P}\left[\widehat{d}_{AB} - d_{AB} > \frac{\alpha_{\text{add}}}{8}\right] &\leq \mathbb{P}\left[\mathcal{E}(\delta)\right] + \mathbb{E}\left[\mathbb{P}\left[\mathcal{E}_2(\delta_2)\middle|\mathcal{E}(\delta)^c, \{\delta_{AB}\}^{(i)}\right]\right] \\
&\leq e^{-m\delta^2} + e^{-mk\delta_2^2}. \qquad (12)
\end{aligned}$$

Substituting the value of δ and δ_2 gives us the necessary result. ■