THE WEIGHTED ENSEMBLE METHOD FOR SAMPLING STEADY STATES

Joint work with:

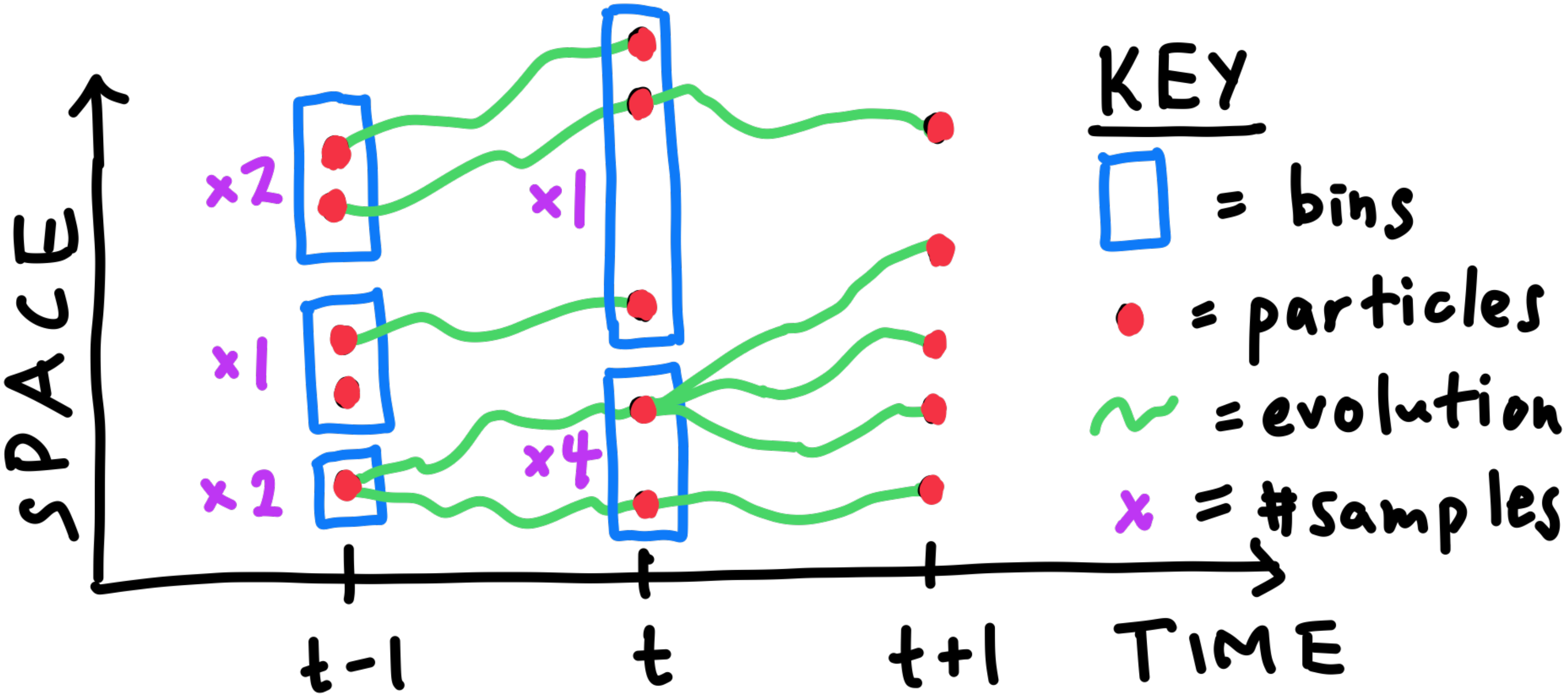
Gideon Simpson (Drexel University)
Rob Webber (NYU/Caltech)
Dan Zuckerman (Oregon Health & Science University)

Special thanks to Mathias Rousset

WEIGHTED ENSEMBLE is an interacting particle importance sampling method.

It is used to estimate distributions of a Markov chain.

The particles evolve according to this Markov chain in between *resampling* steps, where the particles are grouped into *bins* and each bin is resampled.



Why weighted ensemble is important:

1. It is the only importance sampling method for general Markov chain steady states — that admit no explicit formulae — *without a finite particle number bias.*

2. It is a very *simple and flexible* algorithm. You can put particles wherever you like by appropriately choosing bins and the number of “samples” in each bin.

3. With appropriate parameter choices, it approaches the *smallest possible variance* among all unbiased (resampling-based) interacting particle methods.

<u>Weighted ensemble is a variance reduction importance sampling method.</u>

However, it is often used in conjunction with the *Hill relation,* which says that

$$\text{mean first passage time from } \rho \text{ to } B = \frac{1}{\mu(B)}$$

for a Markov chain, with steady state μ, that is recycled at ρ upon reaching B.

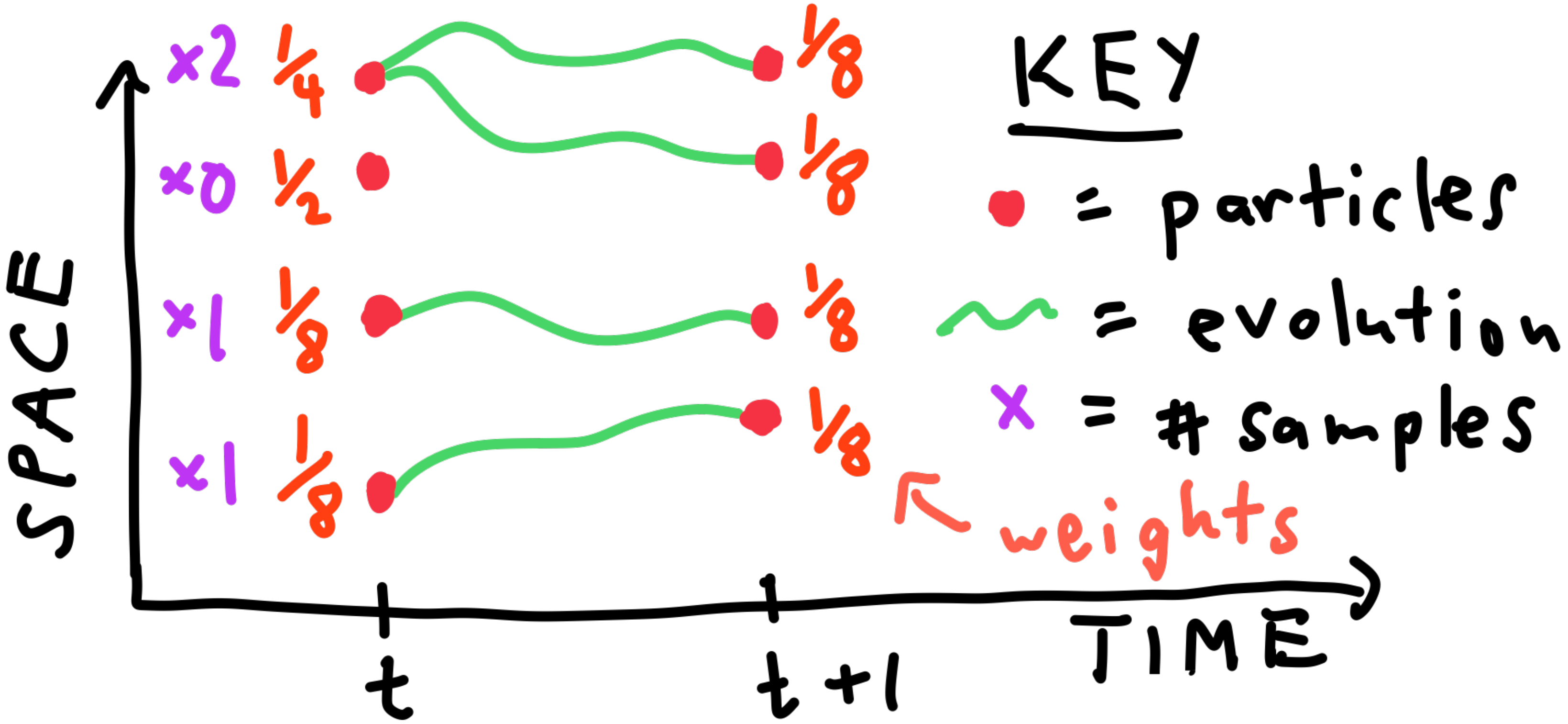
Such mean first passage times are important in many applications — we have in mind computational chemistry, where they could represent the typical time for a ligand (drug) to bind with a protein ($B =$ unbound state). Efficient computation of these times would could pave the way towards in silico drug design.

The Hill relation converts a long-time computation — typically hampered by metastability — to a rare event problem, namely the estimation of $\mu(B) \approx 0$. Introducing the recycling can remove the metastability in many situations.

Unbiased particle methods:
Starting** with weights summing to 1, at each time:

1. Resample from the current particles.
2. Weight of sampled particle = (old weight)/(mean #samples of the particle).
3. Evolve the resampled particles one step according to the Markov chain.

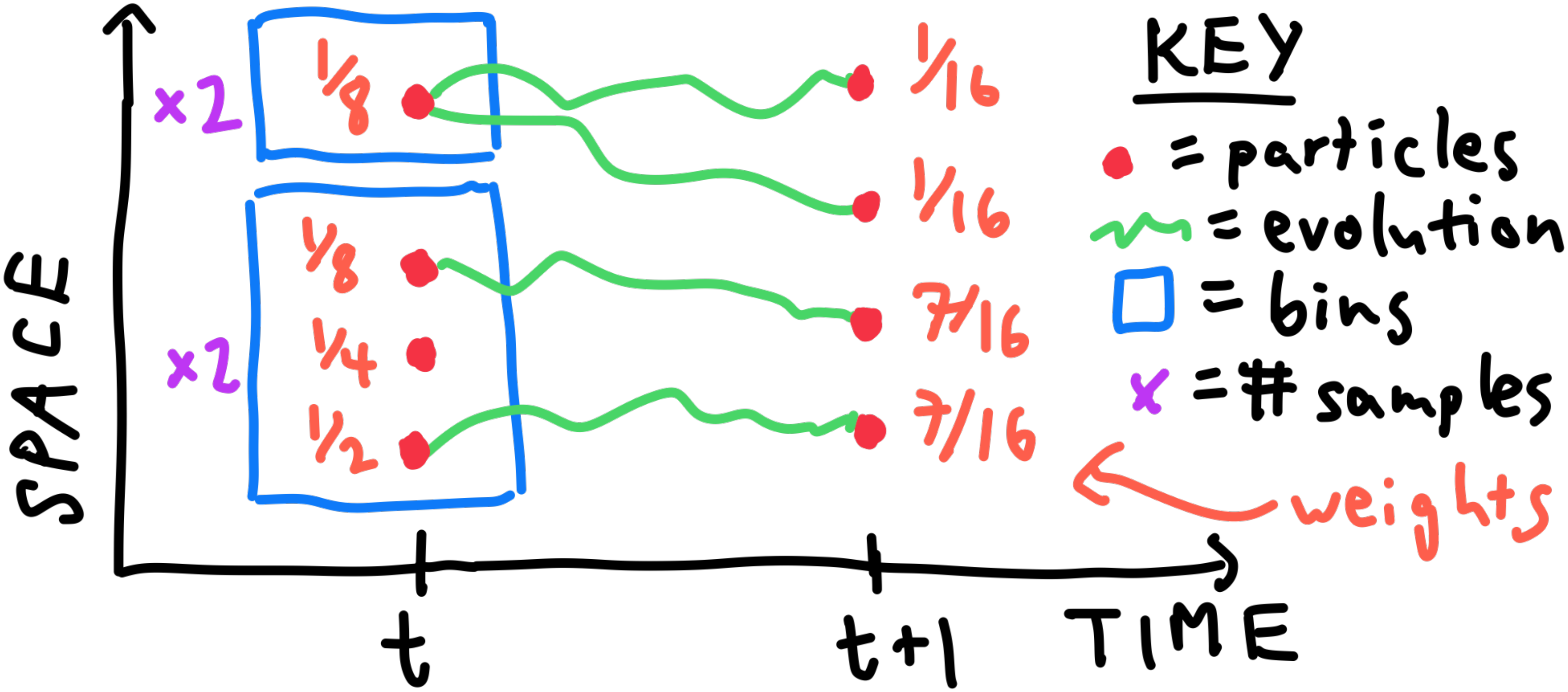
**To simplify presentation, we'll always assume a deterministic initial condition.



Weighted ensemble (special case of unbiased methods):

Starting with weights summing to 1, at each time:

1. Divide (partition) the current particles into bins.
2. In each bin, resample from the particles according to their weights.
3. Assign weight to sampled particle = (its bin weight)/(#samples in its bin).
4. Evolve the resampled particles one step according to the Markov chain.



Before resampling at time t, particles are $\xi_t^1, \ldots, \xi_t^N$ and weights are $w_t^1, \ldots, w_t^N$.

After resampling at time t, particles are $\hat{\xi}_t^1, \ldots, \hat{\xi}_t^N$ and weights are $\hat{w}_t^1, \ldots, \hat{w}_t^N$.

<u>Weight update rule for unbiased methods:</u>

$$\hat{w}_t^i = \frac{w_t^j}{\text{mean \#copies of } \xi_t^j}, \qquad \text{if } \hat{\xi}_t^i \text{ is copied from } \xi_t^j.$$

<u>Weight update rule for weighted ensemble (special case of above):</u>

$$\hat{w}_t^i = \frac{\text{total weight in bin } u}{\text{\#copies in bin } u}, \qquad \text{if } \hat{\xi}_t^i \text{ is copied from a particle in bin } u.$$

Mathematical analysis based on martingale variance decomposition:

The variance of the Doob martingale

$$D_t = \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{N} w_t^i f(\xi_t^i)\,\middle|\,\mathscr{F}_t\right], \hat{D}_t = \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{N} w_t^i f(\xi_t^i)\,\middle|\,\hat{\mathscr{F}}_t\right]$$

decomposes as

$$\text{Var}\left(\underbrace{\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{N} w_t^i f(\xi_t^i)}_{=D_{T-1}}\right) = \underbrace{\frac{1}{T^2}\sum_{t=0}^{T-2}\text{Var}\left(D_{t+1}\,\middle|\,\hat{\mathscr{F}}_t\right)}_{\text{evolution variance}} + \underbrace{\frac{1}{T^2}\sum_{t=0}^{T-2}\text{Var}\left(\hat{D}_t\,\middle|\,\mathscr{F}_t\right)}_{\text{resampling variance}}$$

We’ll assume that f is a bounded function.

Theorem 1. Suppose K, the evolution kernel, is geometrically ergodic wrt μ.

If the weights always sum to 1, then $\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{N} w_t^i f(\xi_t^i) = \mu(f)$ a.s.

If the sum of the weights fluctuates, then this ergodic theorem fails.

Proof idea: Define $h_{t,T}(\xi) = \sum_{s=0}^{T-t-1} K^s f$. By martingale variance decomposition,

$$\mathrm{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{N} w_t^i f(\xi_t^i)\right) =$$

$$\frac{1}{T^2}\sum_{t=0}^{T-2}\mathbb{E}\left[\mathrm{Var}\left(\sum_{i=1}^{N} w_{t+1}^i h_{t+1,T}(\xi_{t+1}^i)\middle|\, \hat{\mathscr{F}}_t\right) + \mathrm{Var}\left(\sum_{i=1}^{N} \hat{w}_t^i K h_{t+1,T}(\hat{\xi}_t^i)\middle|\, \mathscr{F}_t\right)\right].$$

Under *uniform* ergodicity, $h_{t,T} - (T-t)\mu(f)$ is uniformly bounded in t and T.

Thus, *if the weights sum to 1,* then the variance above is $O(1/T)$ as $T \to \infty$.

But, *if the total weight fluctuates at each time,* the variance is “typically” $O(T)$.

Theorem 2:

Any conditionally independent resampling scheme** with weights summing to 1, $w_t^1 + \ldots + w_t^N = \hat{w}_t^1 + \ldots + \hat{w}_t^N = 1$, is equivalent to weighted ensemble.

Proof:

Using $\sum_i \hat{w}_t^i = 1$ and independence, $0 =$ Var $\left(\sum_i \hat{w}_t^i \right) = \sum_i$ Var $\left(\hat{w}_t^i \right)$.

So each $\hat{w}_t^i$ is constant. Choose bins based on the constant values $\hat{w}_t^i$ takes.

**conditionally independent means independent given some auxiliary information, and includes most common methods like multinomial, residual multinomial, stratified, Bernoulli, etc.

<u>Corollary:</u> WE is the <u>ONLY</u> method that converges exactly to $\mu(f)$ w.p. 1.

Other unbiased methods have exploding variance. This can be controlled by dividing by the total weight, but at the cost of a $1/N$ finite particle number bias.

<u>Why would a 1/N bias matter?</u>

Because for ergodic averages, the variance is order $1/(NT)$.
So for large times T, a $1/N$ bias would dominate the mean squared error.***

This differs from typical SMC — based on time marginals — where the bias and variance are both $1/N$, and the variance dominates the mean squared error.

***moreover, in complex problems, typically only small N can be afforded.

Theorem 3: Weighted ensemble can approach the lowest possible variance.

More specifically, assume K is geometrically ergodic wrt μ and define

$w_t(u) =$ total weight in bin u at time t,
$N_t(u) =$ #of samples in bin u at time t,
$\eta_t^u = \sum_{\xi_t^i \in \text{bin } u} \frac{w_t^i}{w_t(u)} \delta_{\xi_t^i} =$ particle distribution in bin u at time t,
$h = \lim_{T \to \infty} (h_{t,T} - (T - t)\mu(f)) =$ solution to Poisson eqn $(I - K)h = f - \mu(f)$.

Then with enough particles and bins and multinomial resampling, if we choose

$$N_t(u) \approx \frac{N w_t(u)\sqrt{\eta_t^u(\text{Var}_K h)}}{\sum_{\text{bins } u} w_t(u)\sqrt{\eta_t^u(\text{Var}_K h)}},$$

then we approach the lowest possible variance among all unbiased methods.

Proof sketch:

$$\text{Var}\left(\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{n} w_t^i f(\xi_t^i)\right) = \text{(using multinomial resampling)}$$

$$\underbrace{\frac{1}{T^2}\sum_{t=0}^{T-2}\mathbb{E}\left[\sum_{\text{bins } u}\frac{w_t(u)^2}{N_t(u)}\eta_t^u(\text{Var}_K h_{t+1,T})\right]}_{\text{evolution variance}} + \underbrace{\frac{1}{T^2}\sum_{t=0}^{T-2}\mathbb{E}\left[\sum_{\text{bins } u}\frac{w_t(u)^2}{N_t(u)}\text{Var}_{\eta_t^u}(K h_{t+1,T})\right]}_{\text{resampling variance}}$$

Jensen’s inequality, unbiasedness, and ergodicity show that

$$\liminf_{T\to\infty} T \times \begin{matrix}\text{evolution}\\ \text{variance}\end{matrix} \geq \frac{1}{N}\left(\int\sqrt{\text{Var}_K h}\, d\mu\right)^2.$$

Using $N_t(u) \approx \propto w_t(u)\sqrt{\eta_t^u(\text{Var}_K h)}$ and enough particles/bins, we approach the RHS in the above inequality, which is thus the smallest possible variance**.

<u>Corollary:</u>

The weighted ensemble variance can approach $\frac{1}{NT}\left(\int\sqrt{\text{Var}_K h}\,d\mu\right)^2$, while the direct Monte Carlo variance is $\approx \frac{1}{NT}\int \text{Var}_K h\,d\mu$.

So, the potential gain depends on how “flat” $\text{Var}_K h$ is.

For instance, if K is an independence sampler, $\text{Var}_K h = 0$ and there is no gain.

Gains occur when there is “incremental progress” towards “important regions.”

In applications, the variance reduction can be many orders of magnitude!

Results on computing tails of the magnetization in the Ising model:

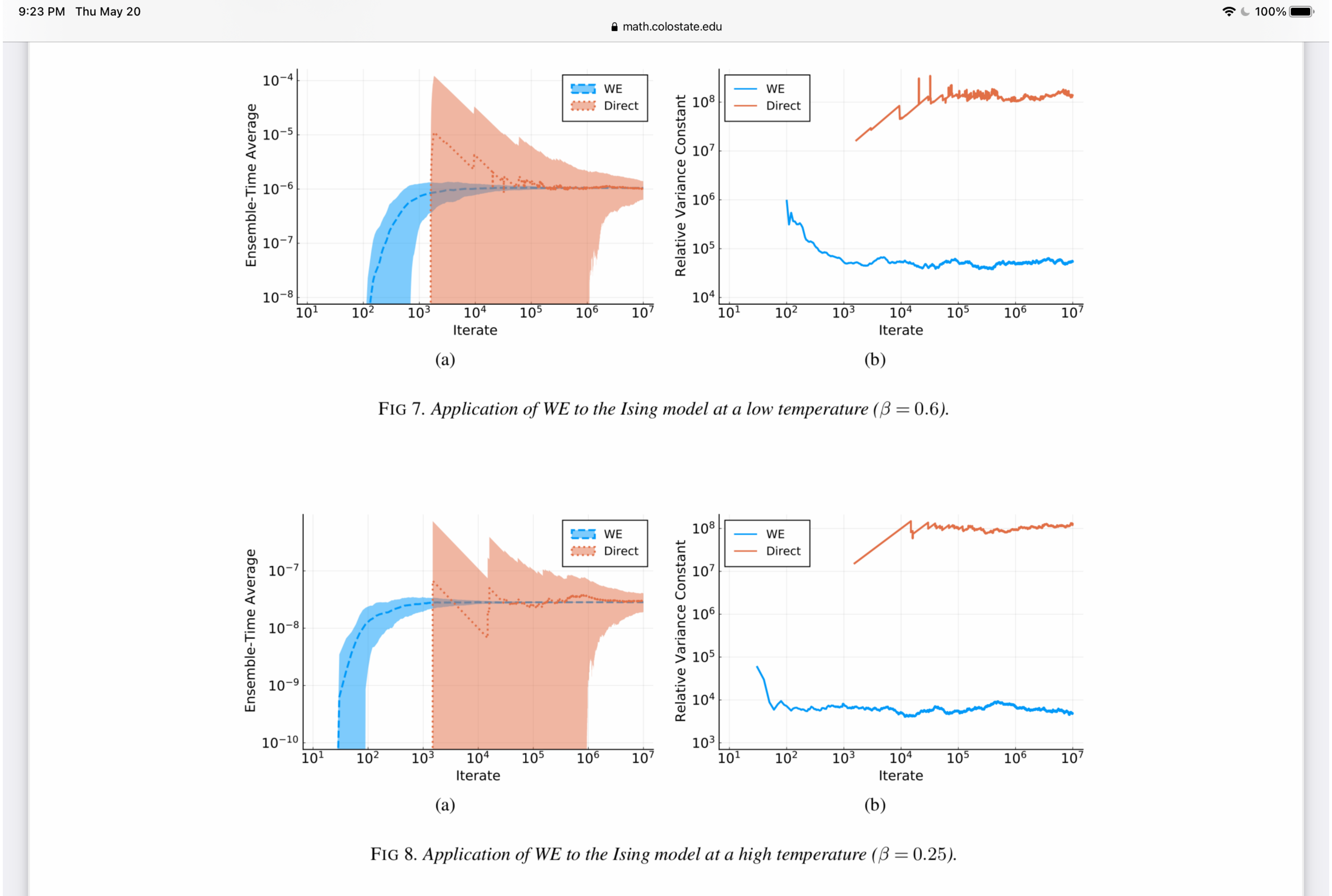


FIG 7. *Application of WE to the Ising model at a low temperature ($\beta = 0.6$).*

FIG 8. *Application of WE to the Ising model at a high temperature ($\beta = 0.25$).*

<u>THANKS TO THE ORGANIZERS FOR THE INVITATION!!!</u>