Importance Sampling

Problem: Estimate

$$E(\mathcal{O}) = \int \mathrm{d}x \; p(x)\mathcal{O}(x)$$

when $\mathcal{O}$ is only non-zero where $p(x)$ is very small.

What is the problem? Suppose

$$\int \mathrm{d}x \; p(x) \; \mathcal{I}_{\mathcal{O}>0} < 10^{-8}.$$

Then for numerical simulations, we have to draw $\approx 10^8$ samples from $p(x)$ to get even one sample where the expectation is non-zero.

Suppose

$$\gamma \equiv \int \mathrm{d}x \; p(x) \; \mathcal{I}_{\mathcal{O}}.$$

Draw N samples from $p(x)$. Then the estimate

$$\gamma_N = \frac{1}{N} \sum_i \mathcal{I}_{\mathcal{O}}(x_i)$$

has variance

$$Var(\gamma_N) = \frac{\gamma(1-\gamma)}{N}.$$

Confidence interval for γ estimate (CLT!):

$$\gamma_N \pm z_{\alpha/2}\sqrt{Var(\gamma_N)}.$$

For example, if $\alpha = 0.01$, the confidence interval for $\gamma < 0.1$ is

$$2.576 \times \sqrt{Var(\gamma_N)} < 0.1\gamma_N.$$

How big must N be?

$$N \approx (\frac{2.576}{0.1})^2 \times \sqrt{\frac{1-\gamma}{\gamma}}.$$

If $\gamma \approx 10^{-6}$ we need almost 7×10^8 samples!

Basic idea of importance sampling

$$\gamma \equiv \int \mathrm{d}x\ p(x)\ \mathcal{I}_{\mathcal{O}} = \int \mathrm{d}x\ q(x)\ \frac{p(x)}{q(x)}\ \mathcal{I}_{\mathcal{O}} = E_q\big(\frac{p(x)}{q(x)}\ \mathcal{I}_{\mathcal{O}}\big).$$

Pick $q(x)$ so that $\mathcal{I}_{\mathcal{O}}(x) = 1$ is *not* rare in the distribution $q(x)$.

Evaluate the new observable $\mathcal{O}_q \equiv \frac{p(x)}{q(x)}\ \mathcal{I}_{\mathcal{O}}$.

Intuitively, maximize $q(x)$ where $\mathcal{I}_{\mathcal{O}}(x) > 0$.

Any q leads to the same expectation with more draws.
So pick q to minimize the variance of $\mathcal{O}_q$.

$$\gamma \equiv \int \mathrm{d}x \; p(x) \, \mathcal{I}_{\mathcal{O}} = \int \mathrm{d}x \; q(x) \, \frac{p(x)}{q(x)} \, \mathcal{I}_{\mathcal{O}} = E_q\big(\frac{p(x)}{q(x)} \, \mathcal{I}_{\mathcal{O}}\big).$$

Pick $q(x)$ so that $\mathcal{I}_{\mathcal{O}}(x) = 1$ is *not* rare in the distribution $q(x)$.

Evaluate the new observable $\mathcal{O}_q \equiv \frac{p(x)}{q(x)} \, \mathcal{I}_{\mathcal{O}}$.

$$E_q\left(\mathcal{O}_q^2\right) = \int \mathrm{d}x \; q(x) \left(\frac{p(x)}{q(x)}\right)^2 \mathcal{I}_{\mathcal{O}}(x) = E_p(\mathcal{O}_q).$$

Does this always work?

NO!

$$p(x) = \lambda \exp(-\lambda x), \quad x \geq 0.$$

We want to estimate $\mathcal{I}_{x>y}$ for some large y.

Try

$$q(x) = \mu \exp(-\mu x)??$$

Then

$$E_p\left(\mathcal{I}_{x>y}\frac{p}{q}\right) = \frac{\lambda^2}{\mu}\int_y^\infty \mathrm{d}x \;\; \exp(-(2\lambda - \mu)x)$$

$$= \frac{\lambda^2}{\mu(2\lambda - \mu)}\left(\exp(-(2\lambda - \mu)y - \exp(-(2\lambda - \mu)L)\right).$$

No minimum at any finite value of μ.

Now, let's apply this IS measure redefinition to actually prove Cramér's theorem. Remember, I didn't actually prove Gärtner-Ellis, I just motivated it, and then cheated by using GE to prove Cramér.

Cramér's Theorem, revisited

Remember Chebyshev. Redo the proof:

$$P(\frac{1}{n}\sum_i X_i > \mu+\epsilon) = P(\sum_i X_i - n\mu > n\epsilon) \geq \frac{E((\sum_i X_i - n\mu)^2)}{n^2\epsilon^2} = \frac{\sigma^2}{n\epsilon^2},$$

and the same argument holds for $P(-\frac{1}{n}\sum_i X_i > \mu+\epsilon)$.

Now, the same argument applies if we take ψ to be any non-negative increasing function:

$$P(\frac{1}{n}\sum_i X_i > s) \leq P(\psi(\sum_i X_i) \geq \psi(ns)) \leq \frac{1}{\psi(ns)}E(\psi(\sum_i X_i)).$$

Take $\psi(s) \equiv \exp(\lambda s)$. Then

$$\frac{1}{n}P(\frac{1}{n}\sum_i X_i > s) \leq -\lambda s + \frac{1}{n}\sum_i \ln E(\exp(\lambda X_i)).$$

Recognize the cumulant generating function? So

$$-\frac{1}{n}P(\frac{1}{n}\sum_i X_i > s) \geq \sup(\lambda s - K_X(\lambda)).$$

Cramér's Theorem: Lower bound

Define a new probability measure

$$q(X) \equiv p(X) \frac{\exp(\lambda X)}{M_X(\lambda)}.$$

This is still normalized:

$$E_q(1) = \frac{E_p(\exp(\lambda X))}{M_X(\lambda)} = 1.$$

Now

$$\frac{1}{n} \ln P(\frac{1}{n} \sum X_i \geq s) \geq \frac{1}{n} \ln P(s + \epsilon > \frac{1}{n} \sum X_i \geq s)$$

$$= \frac{1}{n} \ln E_q \Bigg(\exp(-n\lambda X) M_X^n(\lambda) \mathcal{I}(s + \epsilon > \frac{1}{n} \sum X_i \geq s) \Bigg)$$

$$\geq -\lambda(s + \epsilon) + K_X(\lambda) + \frac{1}{n} \ln P_q(s + \epsilon > \frac{1}{n} \sum X_i \geq s).$$

How do we choose λ :

$$\lim_{\epsilon \downarrow 0} \ln P_q(s+\epsilon > \frac{1}{n} \sum X_i \geq s) \rightarrow 0?$$

Note $K'_X(\lambda) = E_q(X) = E(X \exp \lambda X) K_X(\lambda)^{-1}$,

$$K'_X(\lambda = 0) = \ \mu,$$

and

$$K'_X(\lambda \rightarrow \infty) = \text{maximum of support of } X.$$

So by the mean-value theorem there is some $\lambda(s) \in [0, \infty)$ such that

$$K'_X(\lambda(s)) = s + \frac{\epsilon}{2}.$$

So, finally,

$$-\frac{1}{n} \ln P(\frac{1}{n} \sum X_i \geq s) \leq (\lambda(s)s - K_X(\lambda(s))).$$

Let’s use this probability ‘tilting’ to make ML maximization trivial.

Mayer cluster expansions

For a single spin, taking values in $\{\pm 1\}$, the partition function is

$$Z = \frac{1}{2}\big[\exp(-\epsilon(w+b)) + \exp(-\epsilon(-w+b))\big]$$

where N is the number of configurations and ϵ is the inverse temperature. Write $\exp(-\epsilon V_{ij}) = 1 - (1 - \exp(-\epsilon V_{ij})) \equiv 1 + \Delta_{ij}$. Then

$$Z = \frac{1}{N} \sum_{\text{configs}} \prod_{i<j} (1 + \Delta_{ij}).$$

Expand in Δ :

$$Z = 1 + \frac{1}{N} \sum_{\text{configs}} \Big[\sum_{i<j} \Delta_{ij} + \sum_{i<j} \sum_{k<l} \Delta_{ij} \Delta_{kl} + \dots \Big].$$

In usual statistical mechanics situations, this cluster expansion becomes a more or less geometric expansion because the constituents might be molecules so there are constraints on what configurations can contribute to each separate sum.

For us, things will be much simpler but it's good to know the general idea.

For a single spin, taking values in $\{\pm 1\}$, the partition function is

$$Z = \frac{1}{2}\big[\exp(-\epsilon w) + \exp(-\epsilon(-w))\big]$$

so we get

$$Z = \frac{1}{2}\big[(1 + \Delta_+) + (1 + \Delta_-)\big] = 1 + \frac{1}{2}(\Delta_+ + \Delta_-).$$

For ϵ small we can expand the Δ terms if we want:

$$Z \approx 1 - \frac{\epsilon}{2}(w - w) + \frac{\epsilon^2}{4}(w^2 + (-w)^2) = 1 + \frac{\epsilon^2}{2}(w^2).$$

Take-away message: In this limit Z is trivial to calculate.

For two interacting spins, taking values in $\{\pm 1\}$, the partition function is

$$Z = \frac{1}{4} \sum_{\sigma_1, \sigma_2} \left[\exp(-\epsilon w^i \sigma_i + w^3 \sigma_1 \sigma_2) \right].$$

Now let's be more organized about this expansion. Note (for any variable $\sigma = \pm 1$)

$$\exp(w\sigma) = \cosh w\sigma + \sinh w\sigma = \cosh w + \sigma \sinh w = \cosh w[1 + \sigma \tanh w].$$

So now the cosh terms don't depend on σ values at all so they can come out of the sum over configurations.

$$Z = \frac{1}{4} \cosh(w^1) \cosh(w^2) \cosh(w^3) \sum_{\sigma_1, \sigma_2} (1{+}\sigma_1 \tanh w^1)(1{+}\sigma_2 \tanh w^2)(1{+}\sigma_1 \sigma_2 \tanh w^3).$$

The only terms that will survive in the sum over configurations will be terms that have no linear σ dependence on either of the σ_i, for example, $\tanh w^1 \tanh w^2 \tanh w^3$. For ϵ small, each w comes with a factor of ϵ and $\tanh x \approx x$ for small x. Take-away message: Even with interacting spins, Z is very simple to calculate, especially for small ϵ.

Let's go back to Lecture 1, where we saw

$$\frac{\partial D_{KL}(f||p)}{\partial \theta_i} = E(\mathcal{O}_i) - E_f(\mathcal{O}_i)$$

and we said that

$$E(\mathcal{O}_i) = \partial_{\theta_i} \ln Z(\theta)$$

was difficult to calculate because of Z summing over all configurations.

Was I lying then or am I lying now??

The problem is that we have no idea of the magnitude of the interactions of the spins. In other words, you cannot change the temperature of a system if you have no idea what temperature it's at. All we have is some observed data.

Be concrete: RV: spins $\sigma_i \in \{\pm 1\}, i = 0, \ldots, N-1$.

Spin ‘interactions’: $\mathcal{O}_I \in \{\sigma_i, \sigma_i\sigma_j, \ldots\}$.

Want to assign probabilities for any realization of the RV:

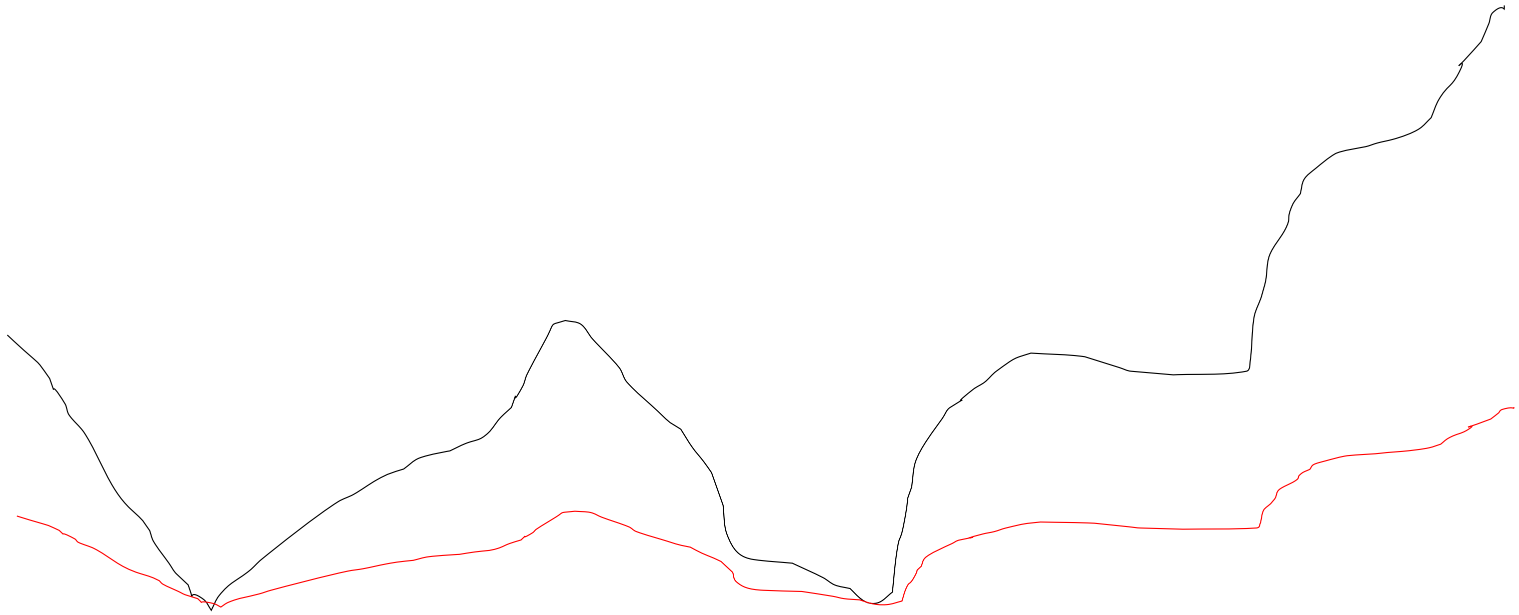
$$P(\sigma|w) \equiv \frac{\exp\left(w^I \mathcal{O}_I(\sigma)\right)}{Z(w)},$$

and ML maximization

$$\Longrightarrow \boxed{w^* : P(\sigma_{\text{obs}}|w^*) = f(\sigma_{\text{obs}}) \equiv \frac{n(\sigma_{\text{obs}})}{\sum_{\text{obs}} n(\sigma_{\text{obs}})}.}$$

$$D_{KL}(f_{\text{obs}}||p) \equiv \sum_{\text{obs}} f_{\text{obs}} \ln\left(\frac{f_{\text{obs}}}{p_{\text{obs}}}\right)$$

Where's the temperature?



Minimizing

$$D_{KL}(f_{\rm obs}||p) \equiv \sum_{\rm obs} f_{\rm obs} \ln\left(\frac{f_{\rm obs}}{p_{\rm obs}}\right)$$

by gradient descent:

$$\delta w^I = \alpha \left[E_f(\mathcal{O}_I) - E_p(\mathcal{O}_I)\right],$$

where (as before)

$$E_q(\mathcal{O}_I) \equiv \sum_{\text{configs } \sigma_i=\pm 1} q(\sigma)\mathcal{O}_I(\sigma).$$

In the small ϵ limit,

$$Z(\epsilon w^I) = 1 + \frac{\epsilon^2}{2} \sum_I (w^I)^2 + \text{h.o. terms.}$$

So in this limit

$$E_{p_\epsilon}(\mathcal{O}_I) = \epsilon w^I,$$

where we define

$$p_\epsilon(\sigma) \equiv \frac{p(\sigma)p(\sigma)^{\epsilon-1}}{\sum_{\text{configs}} p(\text{config})p(\text{config})^{\epsilon-1}}.$$

How is this useful at all? What happened to ensuring that $f_{\rm obs} = p_{\rm obs}$?

Redefine

$$f_\epsilon(\text{obs}) = \frac{f(\text{obs})p(\text{obs})^{\epsilon-1})}{\sum_{\text{all observations}} f(\text{obs})p(\text{obs})^{\epsilon-1})}.$$

Now minimizing

$$D_{KL}(f_\epsilon || p_\epsilon)$$

still gives

$$f(\text{obs}) = p(\text{obs}),$$

but gradient descent is

$$\delta w^I \propto [E_{f_\epsilon}(\mathcal{O}_I) - E_{p_\epsilon}(\mathcal{O}_I)] = \left[E_{f_\epsilon}(\mathcal{O}_I) - \epsilon w^I\right].$$

Why is this useful??

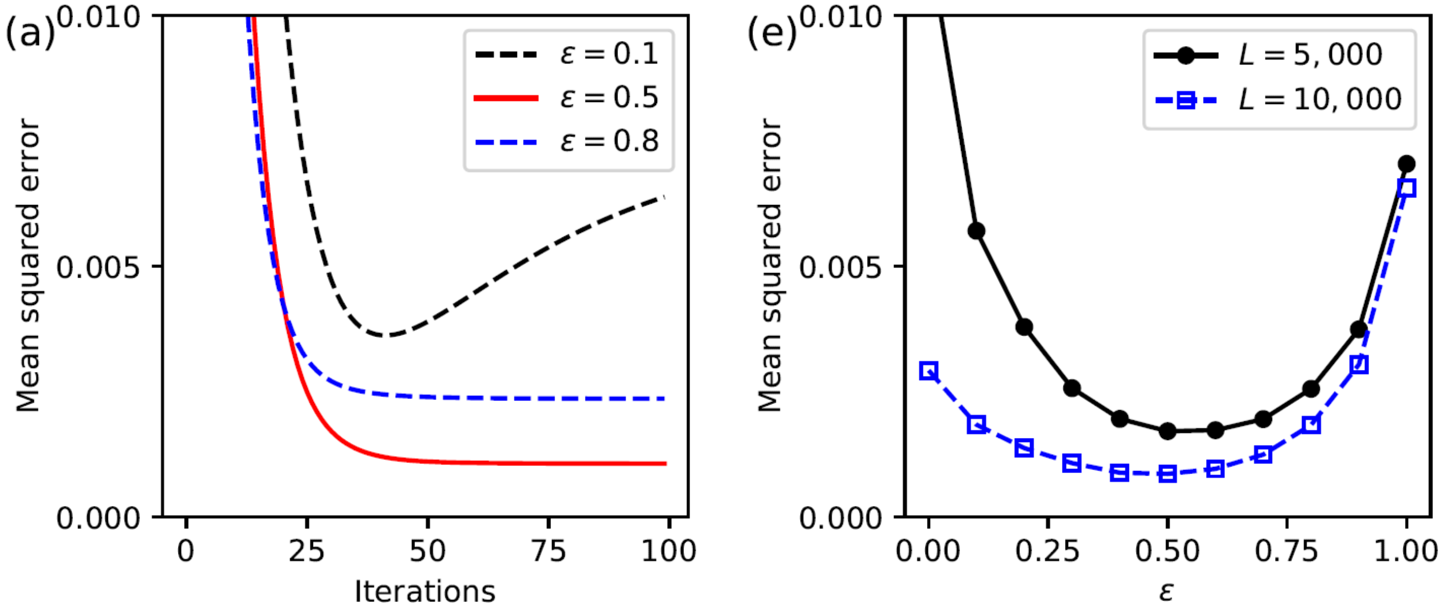
The sum in E_{f_ϵ} is only over observations! NO sum over all configurations.

Hopfield solution: $\epsilon = 1 \implies w^I = E_f(\mathcal{O}_I)$

Analogy with damping term in a differential equation:

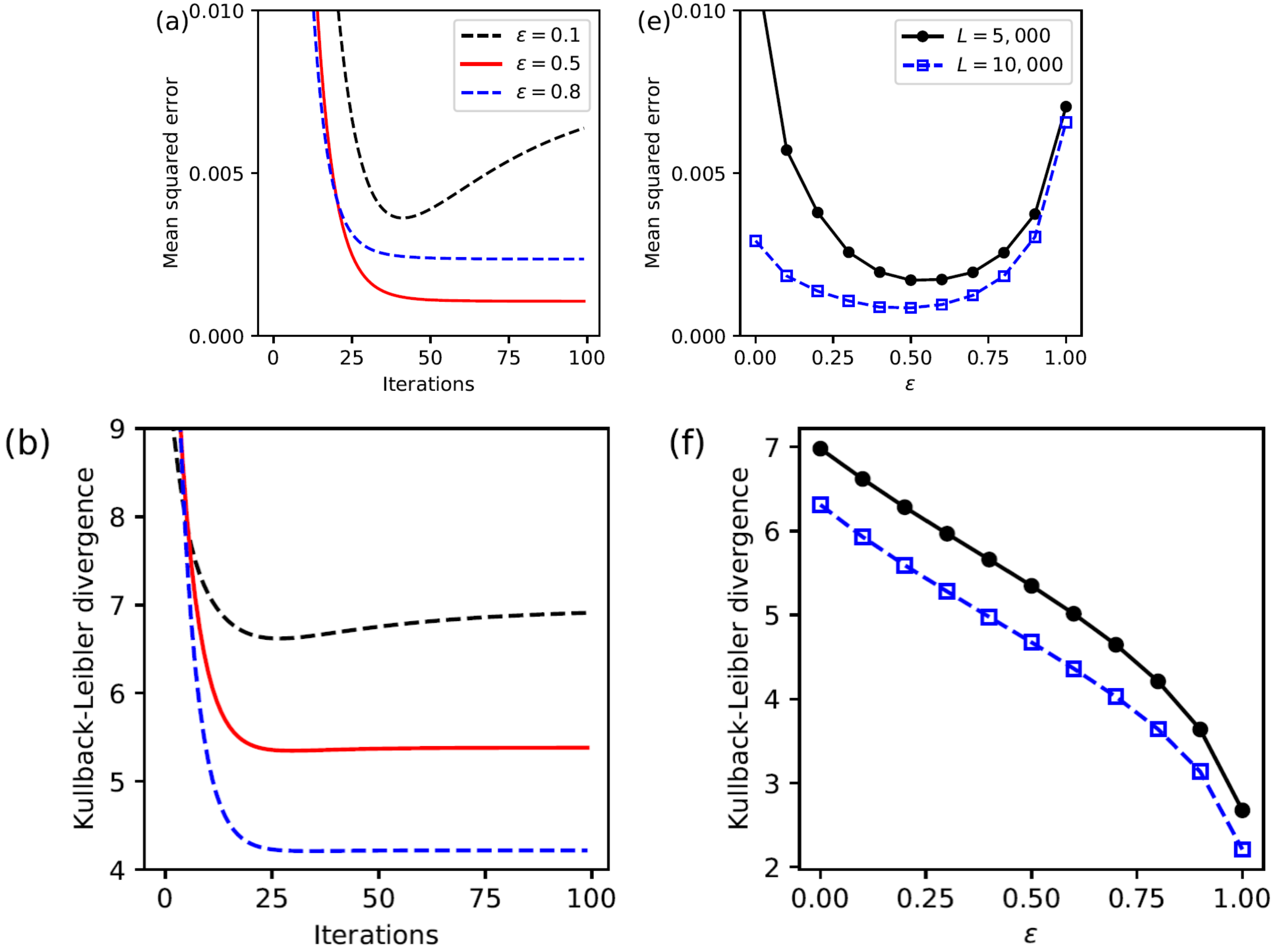
$$\frac{\mathrm{d}w^I}{\mathrm{d}t} = E_{f_\epsilon}(\mathcal{O}_I) - \epsilon w^I .$$

Learning during Iteration

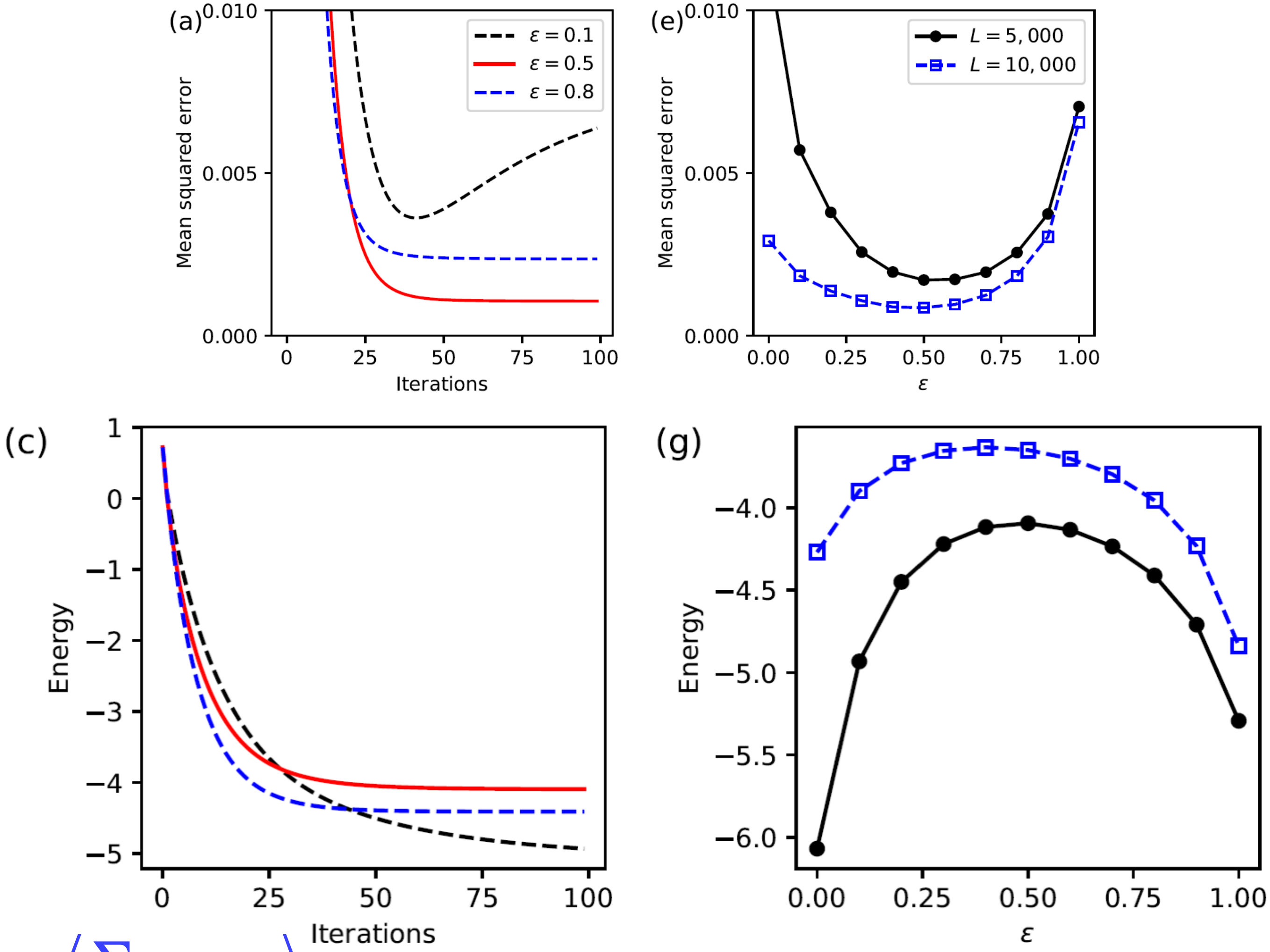


M=20

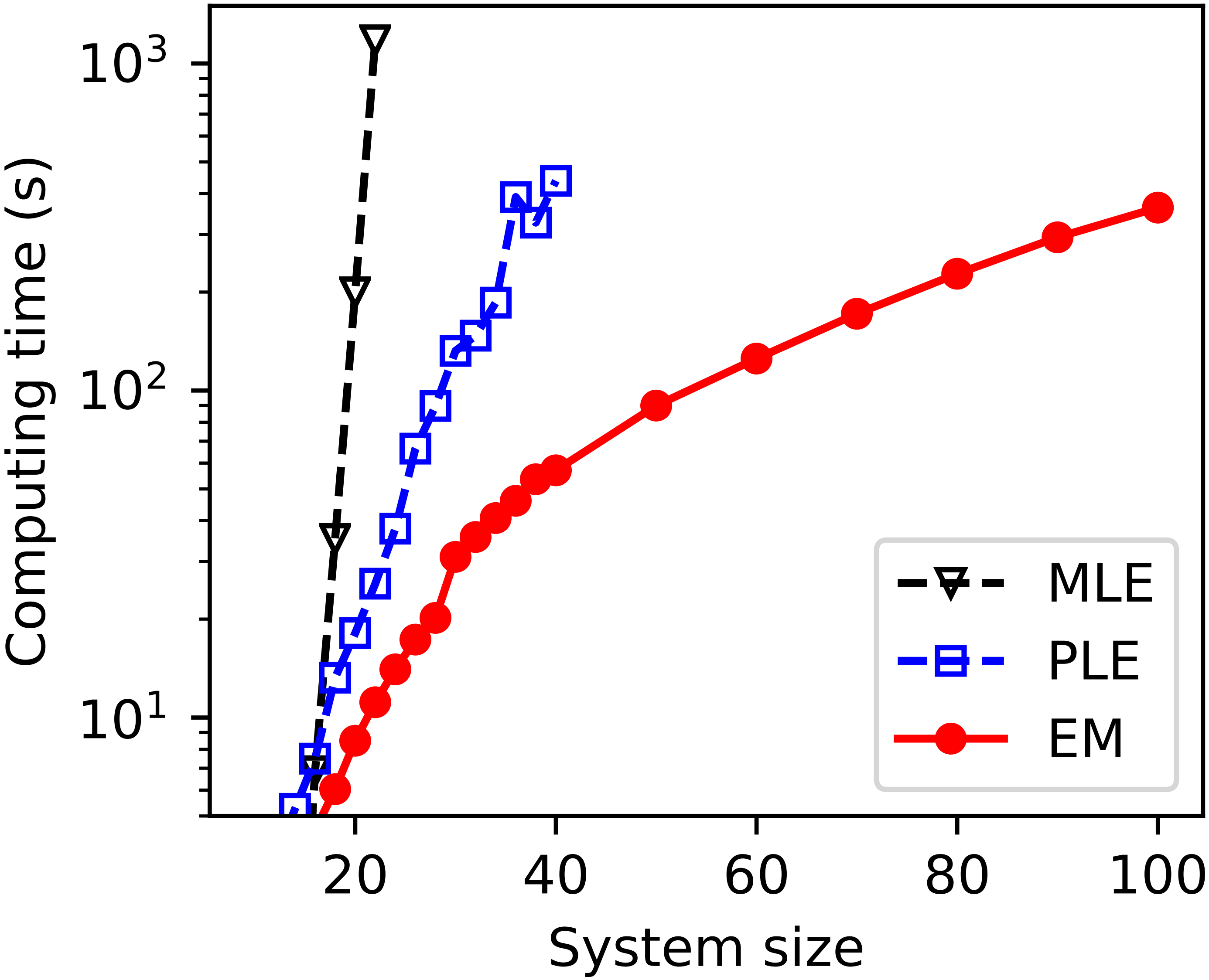
$$MSE = \sqrt{\sum_I \left(w_I - w_I^{true}\right)^2}$$

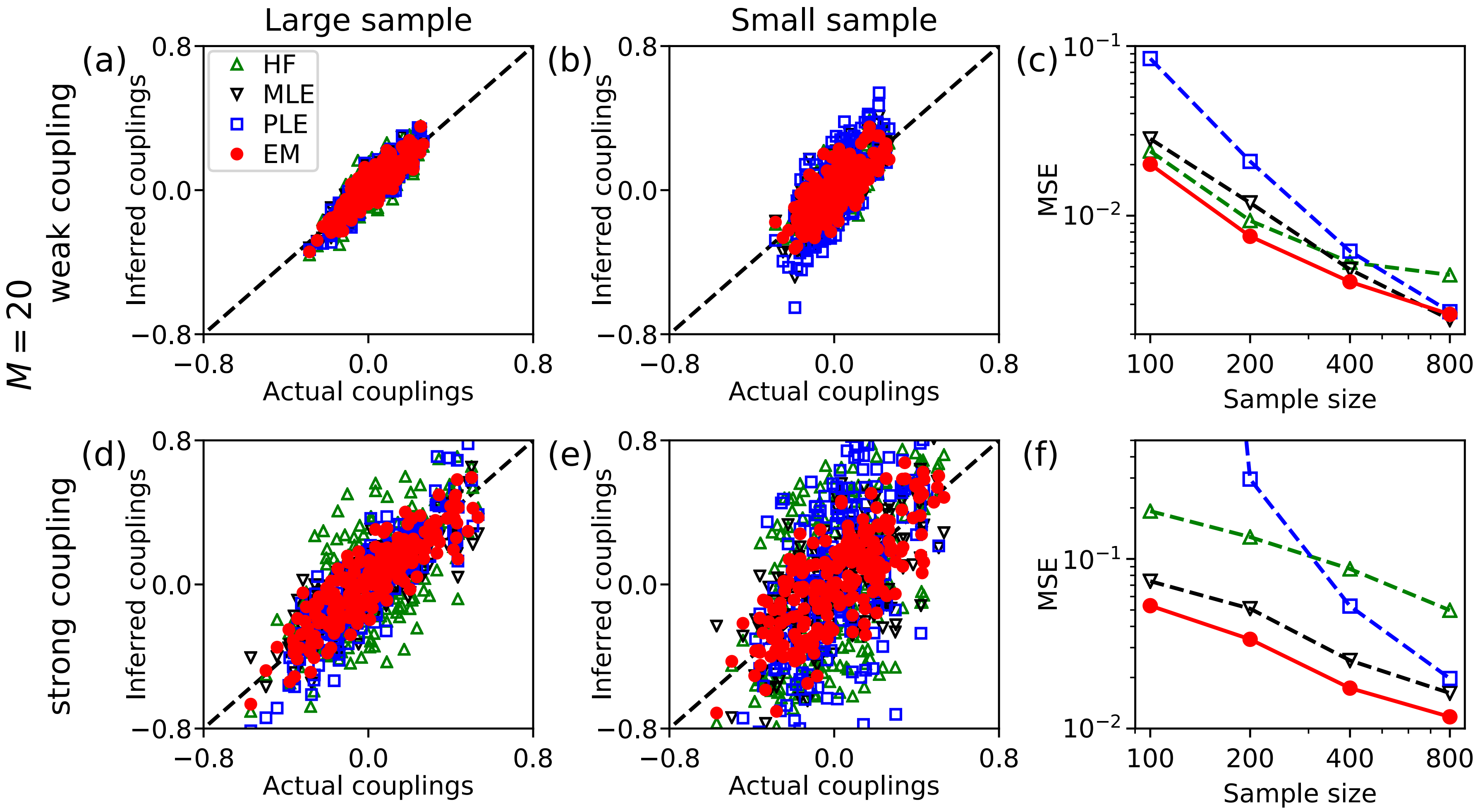


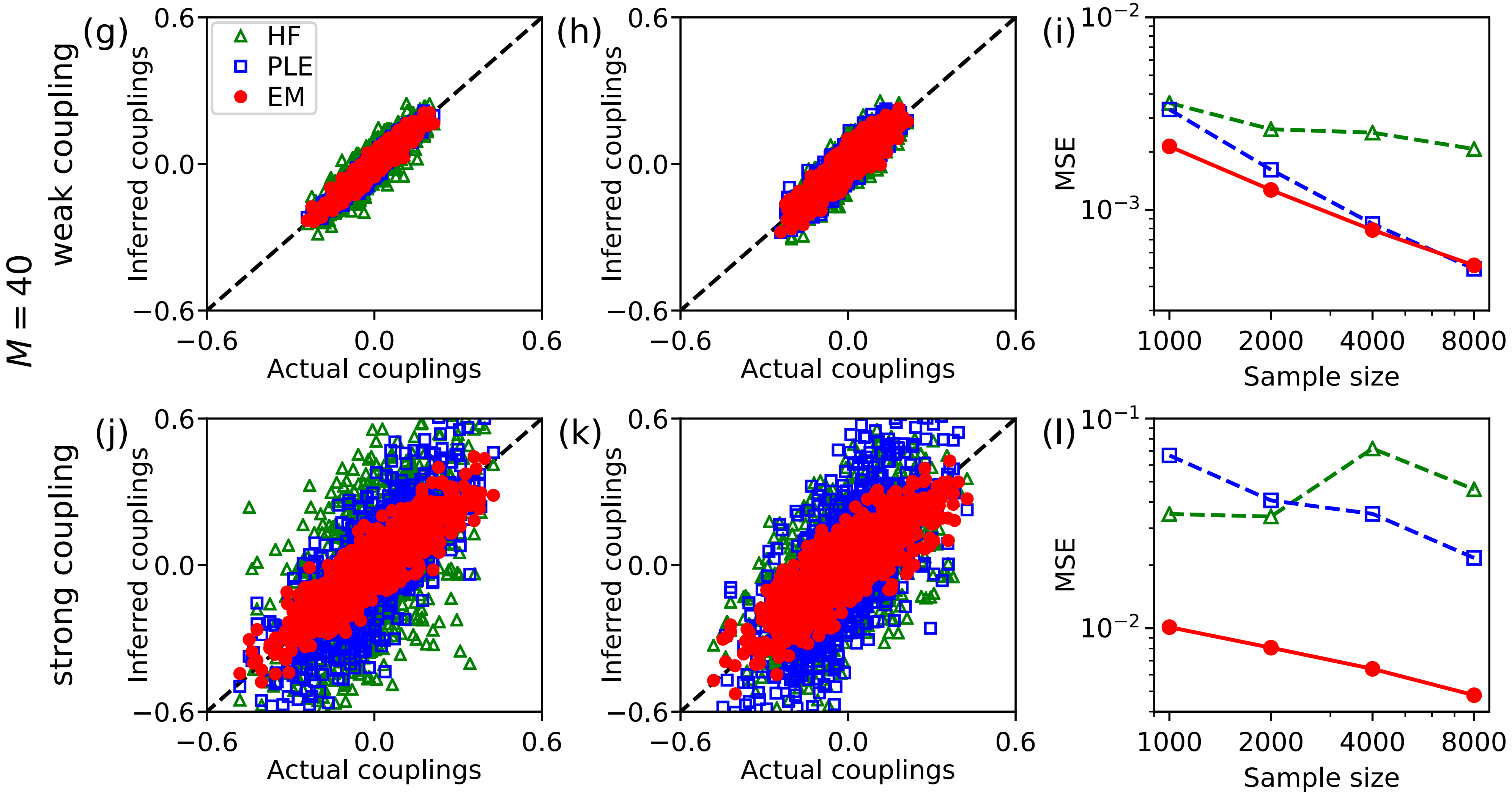
$D_{KL}(\tilde{f}\,||\,\tilde{p})$

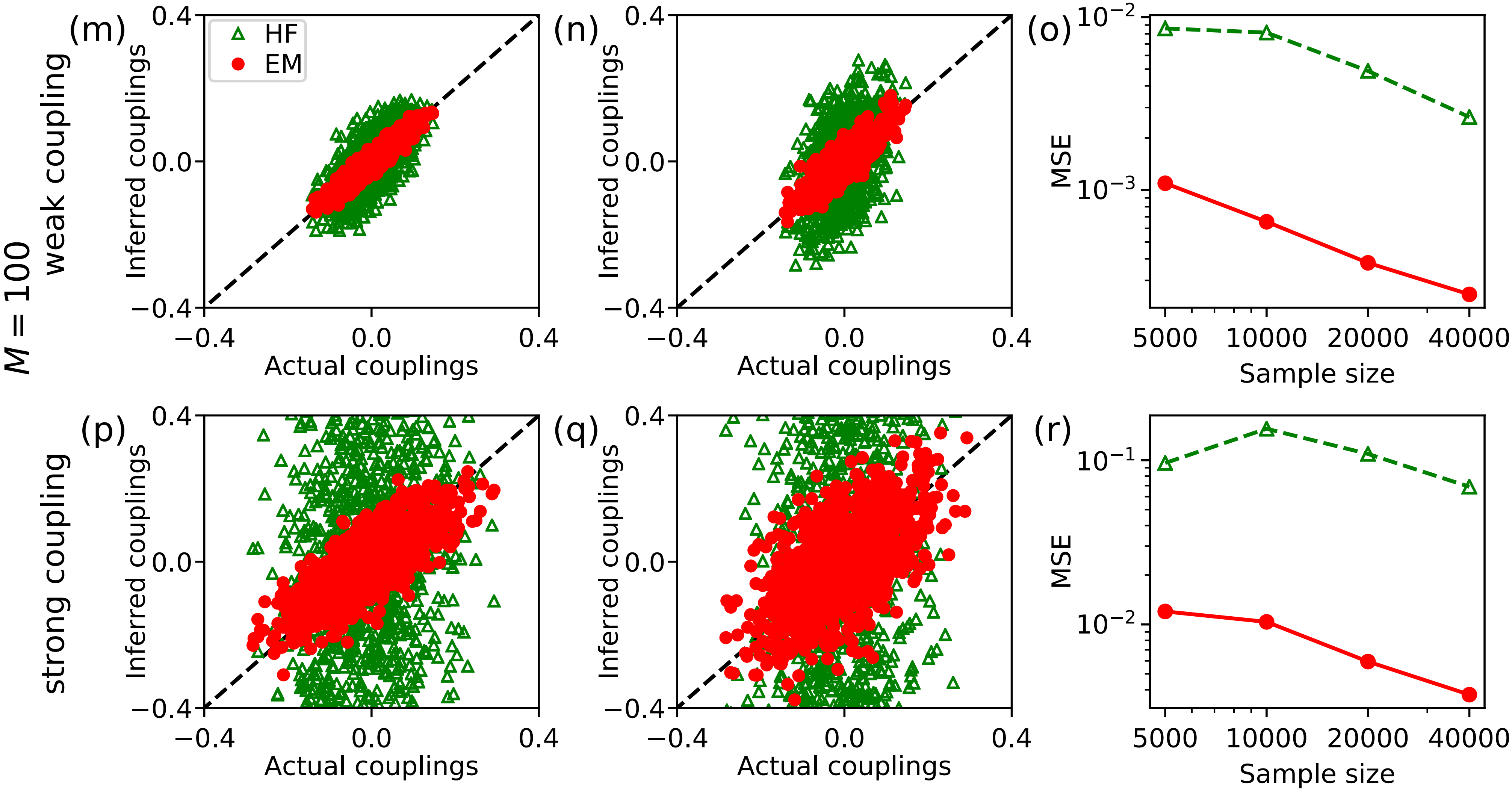


$$\langle E \rangle_f = \left\langle \sum_I w_I O_I(\sigma) \right\rangle_f$$

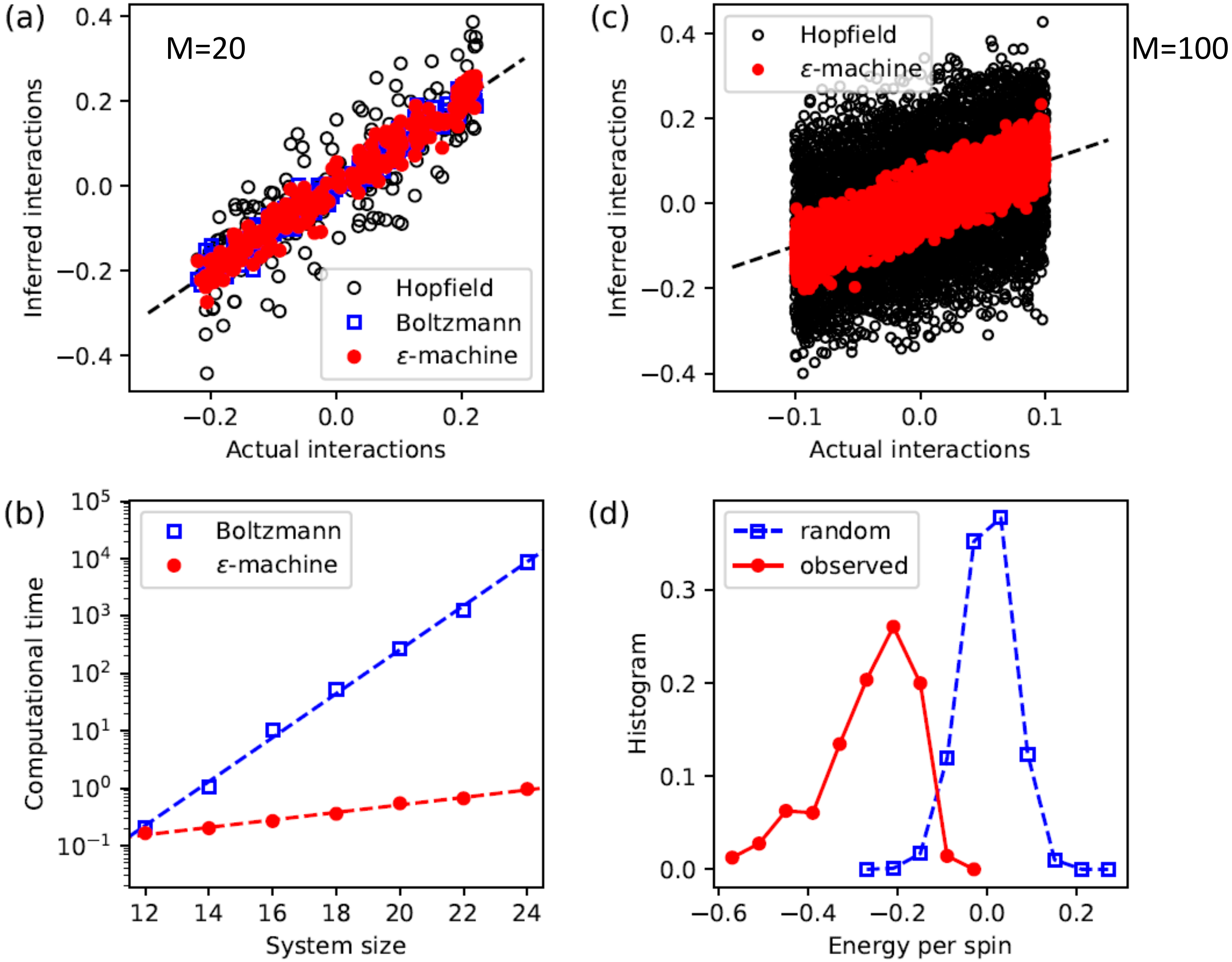

10^3
10^2
10^1
Computing time (s)
20
40
60
80
100
System size
MLE
PLE
EM

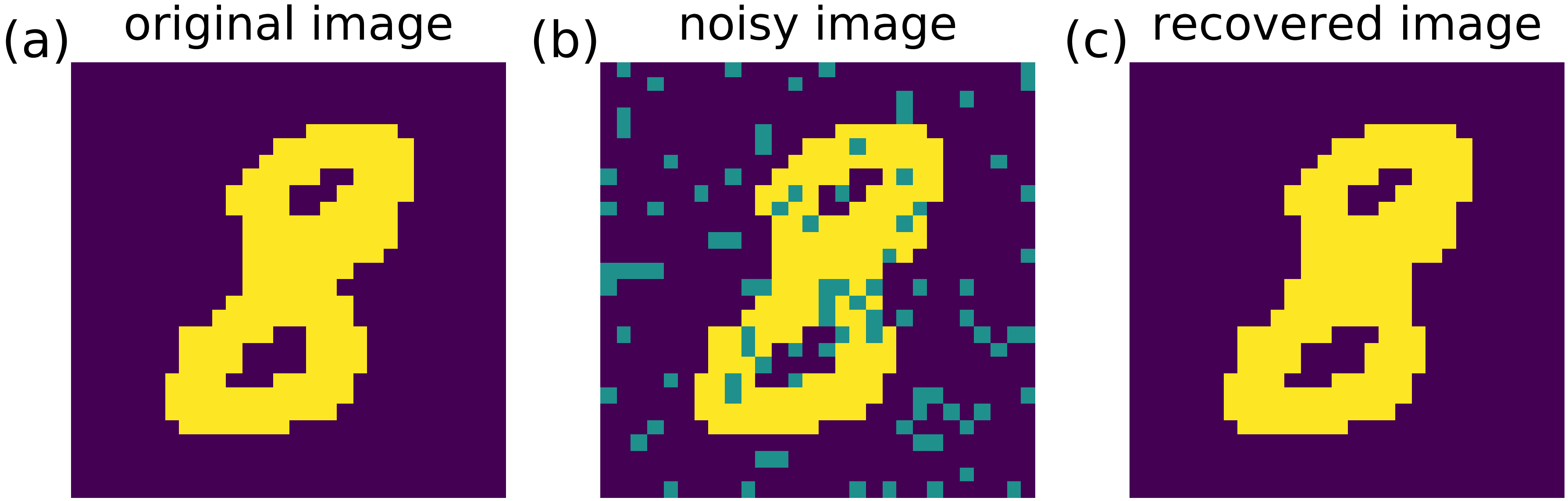

Large sample
Small sample
$M = 20$
weak coupling
strong coupling
(a)
(b)
(c)
(d)
(e)
(f)
HF
MLE
PLE
EM
Inferred couplings
Actual couplings
0.8
0.0
−0.8
MSE
10^{-1}
10^{-2}
100
200
400
800
Sample size


M = 40
weak coupling
strong coupling
(g)
(h)
(i)
(j)
(k)
(l)
HF
PLE
EM
Inferred couplings
Actual couplings
0.6
0.0
−0.6
MSE
Sample size
1000
2000
4000
8000
10^{-2}
10^{-3}
10^{-1}
10^{-2}



$M = 100$
weak coupling
strong coupling
(m)
(n)
(o)
(p)
(q)
(r)
HF
EM
Inferred couplings
Actual couplings
0.4
0.0
−0.4
−0.4
0.0
0.4
MSE
10^{-2}
10^{-3}
10^{-1}
5000
10000
20000
40000
Sample size


(a)
M=20
Inferred interactions
Actual interactions
0.4
0.2
0.0
−0.2
−0.4
−0.2
0.0
0.2
Hopfield
Boltzmann
ε-machine
(c)
M=100
Inferred interactions
Actual interactions
0.4
0.2
0.0
−0.2
−0.4
−0.1
0.0
0.1
Hopfield
ε-machine
(b)
Computational time
System size
10^5
10^4
10^3
10^2
10^1
10^0
10^{-1}
12
14
16
18
20
22
24
Boltzmann
ε-machine
(d)
Histogram
Energy per spin
0.3
0.2
0.1
0.0
−0.6
−0.4
−0.2
0.0
0.2
random
observed


(a)
original image
(b)
noisy image
(c)
recovered image