Appendix

A. Consistency of the Objectives

Theorem 1 (Normalization of solution) *If $\mathbb{E}_p[\tau_t] = 1$, then for any $\lambda > 0$, the estimator (13) has the same solution as (12), hence $\mathbb{E}_p[\tau_{t+1}] = 1$.*

Proof Taking derivative of the objective function in (12) and setting it to zero, we can see that the unconstrained solution is $\frac{\mathcal{T}_p \tau_t}{p}$. Moreover, it satisfies the constraint when $\mathbb{E}_p[\tau_t] = 1$: we can rewrite $\tau_t = \frac{\mu_t}{p}$ for some distribution μ_t and $\mathbb{E}_p\left[\frac{\mathcal{T}_p \tau_t}{p}\right] = \int \mathcal{T}(x'|x)\,\mu_t(x)\,dxdx' = 1$.

We just need to show $\frac{\mathcal{T}_p \tau_t}{p}$ is also the solution to (13). First, note that for any primal τ, the optimal dual v can be attained at $v = \frac{1}{\lambda}(\mathbb{E}_p[\tau] - 1)$. Plugging it to (13), we have for any $\lambda > 0$

$$\begin{aligned}
&\min_{\tau \geqslant 0} \frac{1}{2}\mathbb{E}_{p(x')}\left[\tau^2(x')\right] - \mathbb{E}_{\mathcal{T}_p(x,x')}\left[\tau_t(x)\,\tau(x')\right] + \frac{1}{2\lambda}\left(\mathbb{E}_p[\tau] - 1\right)^2 \\
&\geqslant \min_{\tau \geqslant 0} \frac{1}{2}\mathbb{E}_{p(x')}\left[\tau^2(x')\right] - \mathbb{E}_{\mathcal{T}_p(x,x')}\left[\tau_t(x)\,\tau(x')\right] + \min_{\tau \geqslant 0} \frac{1}{2\lambda}\left(\mathbb{E}_p[\tau] - 1\right)^2 \\
&= -\tfrac{1}{2}\mathbb{E}_p\left[\left(\frac{\mathcal{T}_p \tau_t}{p}\right)^2\right].
\end{aligned} \tag{18}$$

This lower bound is attainable by plugging $\tau = \frac{\mathcal{T}_p \tau_t}{p}$ in (13). Finally, we conclude the proof by noticing that (13) is strictly convex so the optimal solution is unique. ■

B. Convergence Analysis

Let (X, Σ, ν) be a measure space. The $\mathcal{L}^2(X)$ space consists of measurable functions $f : X \mapsto \mathbb{R}$ such that $\|f\| = (\int |f|^2 d\nu)^{1/2} < \infty$. Suppose the initial $\mu_0 \in \mathcal{L}^2(X)$, we want to show the converging behavior of the following damped iteration:

$$\begin{aligned}
\mu_t &= (1 - \alpha_t)\mu_{t-1} + \alpha_t \widehat{\mathcal{T}}\mu_{t-1} \\
&= (1 - \alpha_t)\mu_{t-1} + \alpha_t \mathcal{T}\mu_{t-1} + \alpha_t \epsilon
\end{aligned} \tag{19}$$

with suitable step-sizes $\alpha_t \in (0, 1)$, where $\epsilon \in \mathcal{L}^2(X)$ is a random field due to stochacity in $\widehat{\mathcal{T}}$. To this end, we will use the following lemma.

Lemma 3 *For $\alpha \in \mathbb{R}$, $f, g \in \mathcal{L}^2(X)$*

$$\|(1 - \alpha)f + \alpha g\|^2 = (1 - \alpha)\|f\|^2 + \alpha\|g\|^2 - \alpha(1 - \alpha)\|f - g\|^2.$$

This can be proved by expanding both sides. Now we state our main convergence result.

Theorem 2 *Suppose $\mu_0 \in \mathcal{L}^2(X)$, the step size is $\alpha_t = 1/\sqrt{t}$, $\epsilon \in \mathcal{L}^2(X)$ is a random field and $\mathcal{T}$ has a unique stationary distribution μ. After t iterations, define the probability distribution over the iterations as*

$$\Pr(R = k) = \frac{\alpha_k(1 - \alpha_k)}{\sum_{k'=1}^{t} \alpha_{k'}(1 - \alpha_{k'})}$$

Then there exist some constants $C_1, C_2 > 0$ such that

$$\mathbb{E}\left[\|\mu_R - \mathcal{T}\mu_R\|_2^2\right] \leqslant \frac{C_1}{\sqrt{t}}\|\mu_0 - \mu\|_2^2 + \frac{C_2 \ln t}{\sqrt{t}}\|\epsilon\|_2^2,$$

where the expectation is taken over R. Consequently, μ_R converges to μ for ergodic $\mathcal{T}$.

Proof Using Lemma 3 and the fact that $\mathcal{T}$ is non-expansive, we have

$$\begin{aligned}\|\mu_t - \mu\|^2 &= \|(1-\alpha_t)(\mu_{t-1}-\mu) + \alpha_t(\mathcal{T}\mu_{t-1}-\mu) + \alpha_t\epsilon\|^2 \\ &\leqslant \|(1-\alpha_t)(\mu_{t-1}-\mu) + \alpha_t(\mathcal{T}\mu_{t-1}-\mu)\|^2 + \alpha_t^2\|\epsilon\|^2 \\ &\leqslant (1-\alpha_t)\|\mu_{t-1}-\mu\|^2 + \alpha_t\|\mathcal{T}\mu_{t-1}-\mu\|^2 - \alpha_t(1-\alpha_t)\|\mu_{t-1}-\mathcal{T}\mu_{t-1}\|^2 + \alpha_t^2\|\epsilon\|^2 \\ &\leqslant \|\mu_{t-1}-\mu\|^2 - \alpha_t(1-\alpha_t)\|\mu_{t-1}-\mathcal{T}\mu_{t-1}\|^2 + \alpha_t^2\|\epsilon\|^2.\end{aligned}$$

Then telescoping sum gives

$$0 \leqslant \|\mu_t-\mu\|^2 \leqslant \|\mu_0-\mu\|^2 + \sum_{k=1}^{t}\alpha_k^2\|\epsilon\|^2 - \sum_{k=1}^{t}\alpha_k(1-\alpha_k)\|\mu_k-\mathcal{T}\mu_k\|^2$$

So

$$\sum_{k=1}^{t}\alpha_k(1-\alpha_k)\|\mu_k-\mathcal{T}\mu_k\|^2 \leqslant \|\mu_0-\mu\|^2 + \sum_{k=1}^{t}\alpha_k^2\|\epsilon\|^2.$$

Divide both sides by $\sum_{k=1}^{t}\alpha_k(1-\alpha_k)$ (taking expectation over iterations) gives

$$\mathbb{E}[\|\mu_R-\mathcal{T}\mu_R\|^2] = \sum_{k=1}^{t}\frac{\alpha_k(1-\alpha_k)}{\sum_{k'}\alpha_{k'}(1-\alpha_{k'})}\|\mu_k-\mathcal{T}\mu_k\|^2 \leqslant \frac{\|\mu_0-\mu\|^2 + \sum_{k=1}^{t}\alpha_k^2\|\epsilon\|^2}{\sum_{k=1}^{t}\alpha_k(1-\alpha_k)}.$$

When $\alpha_t = 1/\sqrt{t}$, we have

$$\begin{aligned}\sum_{k=1}^{t}\alpha_k^2\|\epsilon\|^2 &= \sum_{k=1}^{t}\frac{1}{k}\|\epsilon\|^2 \leqslant (\ln t+1)\|\epsilon\|^2 \\ \sum_{k=4}^{t}\alpha_k(1-\alpha_k) &= \sum_{k=4}^{t}\frac{1}{\sqrt{k}} - \frac{1}{k} \geqslant \int_4^t\left(\frac{1}{\sqrt{k+1}} - \frac{1}{k+1}\right)dk = \Omega\left(t^{\frac{1}{2}}\right)\end{aligned}$$

So for big enough t, there exists $C_0 > 0$ such that

$$\mathbb{E}\left[\|\mu_R-\mathcal{T}\mu_R\|^2\right] \leqslant \frac{\|\mu_0-\mu\|^2 + \ln(t+1)\|\epsilon\|^2}{C_0\sqrt{t}},$$

which leads to the the bound in the theorem and $\mathbb{E}\left[\|\mu_R-\mathcal{T}\mu_R\|^2\right] = \widetilde{\mathcal{O}}\left(t^{-1/2}\right)$. Additionally, since $\mathcal{T}$ has a unique stationary distribution $\mu = \mathcal{T}\mu$, we have μ_R converges to μ. ■

C. Application to Off-policy Stationary Ratio Estimation

We provide additional details describing how the variational power method we have developed in the main body of the paper can be applied to the behavior-agnostic off-policy estimation problem (OPE). The general framework has been introduced in Section 1 and the implementation for the undiscounted case ($\gamma = 1$) is demonstrated in Section 5.4. Specifically, given a sample $\mathcal{D} = \left\{(s,a,r,s')_{i=1}^n\right\}$ from the behavior policy, we compose each transition in $\mathcal{D}$ with a target action $a' \sim \pi(\cdot|s')$. Denoting $x = (s,a)$, the data set can be expressed as $\mathcal{D} = \left\{(x,x')_{i=1}^n\right\}$. Applying the proposed VPM with $\mathcal{T}(x'|x)$, we can estimate $\frac{\mu(s,a)}{p(s,a)}$. Here the $\mu(s,a) = d_\pi(s)\pi(a|s)$ consists of the stationary state occupancy d_π and the target policy π, while $p(s,a)$ is the data-collecting distribution. Then the average accumulated reward can be obtained via (3).

Here we elaborate on how the discounted case (*i.e.*, $\gamma \in (0,1)$) can be handled by our method. We first introduce essential quantities similar to the undiscounted setting. For a trajectory generated stochastically using policy π from an initial state s_0: $(s_0,a_0,r_0,s_1,a_1,r_1,\ldots)$, where $a_t \sim \pi(\cdot|s_t)$, $s_{t+1} \sim P(\cdot|s_t,a_t)$ and $r_t \sim R(s_t,a_t)$, the the policy value is

$$\rho_\gamma(\pi) := (1-\gamma)\,\mathbb{E}_{s_0\sim\mu_0,a\sim\pi,s'\sim P}\left[\sum_{t=0}^{\infty}\gamma^t r_t\right],$$

where μ_0 is the initial-state distribution. Denote

$$d_t^\pi(s,a) = \mathbb{P}\left(s_t = s, a_t = a \middle| \begin{bmatrix} s_0 \sim \mu_0, \forall i < t, \\ a_i \sim \pi(\cdot|s_i), \\ s_{i+1} \sim P(\cdot|s_i, a_i) \end{bmatrix} \right).$$

The discounted occupancy distribution is

$$\mu_\gamma(s,a) := (1-\gamma) \sum_{t=0}^{\infty} \gamma^t d_t^\pi(s,a). \tag{20}$$

Then, we can re-express the discounted accumulated reward via μ_γ and the stationary density ratio,

$$\rho_\gamma(\pi) = \mathbb{E}_{(s,a)\sim\mu_\gamma(s,a)}[r(s,a)] = \mathbb{E}_{(s,a)\sim p(s,a)}\left[\frac{\mu_\gamma(s,a)}{p(s,a)} r(s,a) \right]. \tag{21}$$

The proposed VPM is applicable to estimating the density ratio in this discounted case. Denoting $x = (s,a)$, $x' = (s',a')$ respectively for notational consistency, we expand μ_γ and use the definition of d_t^π:

$$\begin{aligned} \mu_\gamma(s',a') &= (1-\gamma)\mu_0(s')\pi(a'|s') + \gamma \int \pi(a'|s') P(s'|s,a) \mu_\gamma(s,a) \, ds \, da \\ \implies p(x')\tau^*(x') &= (1-\gamma)\mu_0\pi(x') + \gamma \int \mathcal{T}_p(x,x')\tau^*(x) \, dx, \end{aligned} \tag{22}$$

where $\mu_0\pi(x') = \mu_0(s')\pi(a'|s')$ and $\mathcal{T}_p(x,x') = \pi(a'|s')P(s'|s,a)p(s,a)$.

It has been shown that the RHS of (22) is contractive (Sutton and Barto, 1998; Mohri et al., 2012), therefore, the fix-point iteration,

$$p(x')\tau_{t+1}(x') = (1-\gamma)\mu_0\pi(x') + \gamma \int \mathcal{T}_p(x,x')\tau_t(x) \, dx, \tag{23}$$

converges to the true τ as $t \to \infty$, provided the update above is carried out exactly. Compared to (6), we can see that the RHS of (23) is now a mixture of $\mu_0\pi$ and $\mathcal{T}_p$, with respective coefficients $(1-\gamma)$ and γ.

Similarly, we construct the $(t+1)$-step variational update as

$$\tau_{t+1} = \arg\min_{\tau \geqslant 0} \tfrac{1}{2}\mathbb{E}_{p(x')}\left[\tau^2(x')\right] - \gamma\mathbb{E}_{\mathcal{T}_p(x,x')}[\tau_t(x)\tau(x')] - (1-\gamma)\mathbb{E}_{\mu_0 p(x')}[\tau(x')] + \lambda\left(\mathbb{E}_p[\tau] - 1\right)^2. \tag{24}$$

Compared to (11), we see that the main difference is the third term of (24) involves the initial distribution. As $\gamma \to 1$, (24) reduces to (11).

D. Experiment Details

Here we provide additional details about the experiments. In all experiments, the regularization $\lambda = 0.5$ and the optimizer is Adam with $\beta_1 = 0.5$. The τ model is a neural network with 2 hidden layers of 64 units each with ReLU activation and softplus activation for the output.

D.1. Queueing

For Geo/Geo/1 queue, when the arrival and finish probabilities are $q_a, q_f \in (0,1)$ respectively with $q_f > q_a$, the stationary distribution is $P(X=i) = (1-\rho)\rho^i$ where $\rho = q_a(1-q_f)/[q_f(1-q_a)]$ (Serfozo, 2009, Sec.1.11). The defaults are $(n, q_a, q_f) = (100, 0.8, 0.9)$ for the figures. ρ is called *traffic intensity* in the queueing literature and we set $B = \lceil 40\rho \rceil$ in the experiment. The mean and standard error of the log KL divergence is computed based on 10 runs. We conduct closed-form update for 1000 steps. As for the model-based method, we simulate the transition chain for 200 steps to attain the estimated stationary distribution.

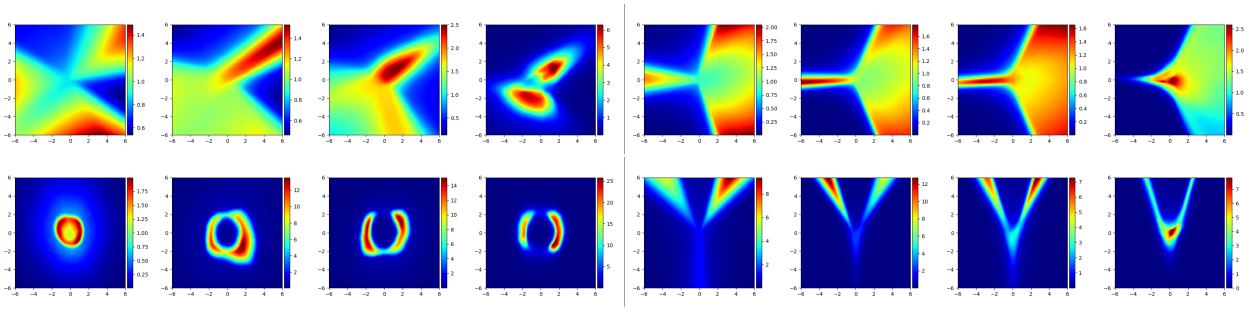

Figure 7. The VPM estimates after $\{10, 20, 30, 150\}$ iterations on the datasets. As we can see, with the algorithm proceeds, the learned stationary density ratio is getting closer to the ground-truth.

D.2. Solving SDEs

Using initial samples are uniformly spaced in $[0, 1]$, we run the Euler-Maruyama (EM) method and evaluate the MMD along the path. The τ model is a neural network with 2 hidden layers of 64 units each with ReLU and Softplus for the final layer. Numbers of outer and inner steps are $T = 50, M = 10$. The learning rate is 0.0005. At each evaluation time step t, we use the most recent 1% of evolution data to train our model τ. The plots are reporting the mean and standard deviation over 10 runs. For the phylogeny studies, the number of particles is $1k$ and $dt = 0.0005$ for the EM simulation, while the rest settings using $dt = 0.001$.

D.3. Post-processing MCMC

The potential functions are collected from several open-source projects[1][2]. $50k$ examples are sampled from the uniform distribution $p(x) = \text{Unif}(x; [-6, 6]^2)$, then transition each x through an HMC operator (one leapfrog step of size 0.5). The model-based $\widehat{\mathcal{T}}$ has a similar structure as τ except the final layer has 2D output without activation to estimate the Gaussian mean. The mini-batch size is $B = 1k$, the maximum number of power iterations $T = 150$ and the number of inner optimization steps is $M = 10$. The model-based $\mathcal{T}$ is given the same number of iterations ($MT = 1500$). The learning rate is 0.001 for τ and 0.0005 for $\widehat{\mathcal{T}}$. To compute the model-based sample, we apply the estimated transition 100 time steps. The MMD plot is based on a "true sample" of size $2k$ from the stationary distribution (estimated by 2k HMC transition steps). The numbers are mean and standard deviation over 10 runs. The MMD is computed by the Gaussian kernel with the median pairwise distance as kernel width.

The quality of the transition kernel and the generated data is critical. Since x and x' are supposed to be related, we use an HMC kernel with one leap-frog step. The initial x is effectively forgotten if using too many leap-frog steps. The main point is to show that our method can utilize the intermediate samples from the chain other than the final point. Moreover, to conform with Assumption 2, the potential functions are numerically truncated.

To verify the convergent behavior of our method, Fig. 7 shows how the ratio network improves as we train the model. It can be seen that the our method quickly concentrates its mass to the region with high potentials.

D.4. Off-policy Evaluation

Taxi is a 5×5 gridworld in which the taxi agent navigates to pick up and drop off passengers in specific locations. It has a total of 2000 states and 6 actions. Each step incurs a -1 reward unless the agent picks up or drops off a passenger in the correct locations. The behavior policy is set to be the policy after 950 Q-learning iterations and the target policy is the policy after 1000 iterations. In the Taxi experiment, given a transition (s, a, s'), instead of sampling one single action from the target policy $\pi(a'|s')$, we use the whole distribution $\pi(\cdot|s')$ for estimation. We conduct closed-form update in the power method and the number of steps is $T = 100$.

Continuous experiments. The environments are using the open-source PyBullet engine. The state spaces are in $\mathbb{R}^9, \mathbb{R}^{26}, \mathbb{R}^{28}$ respectively and the action spaces are in $\mathbb{R}^2, \mathbb{R}^6, \mathbb{R}^8$ respectively. the τ model is the same as in the SDE

[1] https://github.com/kamenbliznashki/normalizing_flows
[2] https://github.com/kevin-w-li/deep-kexpfam

experiment (except for input, which depends on the environment). $T = 200, M = 10, B = 1k$ and the learning rate is 0.0003. The model-based method has a similar neural network structure and is trained for $MT = 2k$ steps with a learning rate of 0.0005. The target policy for the Reacher agent is pretrained using PPO while the HalfCheetah and Ant agents are pretrained using A2C (all with two hidden layers of 64 units each).

The results in the plots are mean and standard deviation from 10 runs.