A NOTATION

n	Total sample
μ	Baseline measure
$p(x;\theta)$	Unnormalized Model
$\tilde{p}(x;\theta)$	Normalized model
c	Normalizing constant parameter
τ	$(c,\, \theta^\top)^\top$
$q(x;\tau)$	One-parameter extended model $\exp(-c)p(x;\theta)$
Θ	Parameter space for θ
Θ_τ	Parameter space for τ
$\eta^*(x)$	True density
$\hat{\eta}_n(x)$	Nonparametric estimator
F_{η^*}	True distribution
p_n	Empirical density
E_*	Expectation under true distribution
$\tilde{\mathrm{E}}$	Expectation under empirical distribution
Var_*	Variance under true distribution
∇_x	Differentiation with respect to x
$\mathbb{P}_n$	Empirical distribution of n samples from F_{η^*}
$\mathbb{G}_n$	Empirical process $\sqrt{n}(\mathbb{P}_n - \mathrm{F}_{\eta^*})$
$\mathfrak{I}_\theta$	Fisher information matrix for θ
$\|_{\tau^*}$	the value at $\tau = \tau^*$
$\mathcal{N}(A, B)$	Normal distribution with mean A, variance B
$L^2(\mathrm{F}_{\eta^*})$	L^2-space with the underlying distribution F_{η^*}
$\mathcal{X}$	Sample space
$B_f(u, v)$	Bregman divergence based on f between u and v
K	Kernel
$\hat{\tau}_{\mathrm{s}}$	Self density-ratio matching estimator with a separable divergence. Note that it is equal to $(\hat{c}_{\mathrm{s}}, \hat{\theta}_{\mathrm{s}})$
$\hat{\tau}_{\text{ns-}\gamma}$	Self density-ratio matching estimator with a γ-divergence
$\hat{\tau}_{\text{ns-ps}}$	Self density-ratio matching estimator with a pseudo spherical divergence
$\|\cdot\|$	Euclidean norm
$\|\cdot\|_\infty$	l_∞ norm

B CONVEXITY

We see specific examples of $f(x)$, satisfying the above conditions in Theorem 5.

Example B.1 *For the functions $f(z) = z\log z$ and $f(z) = 2z\log z - 2(1+z)\log(1+z)$, we can confirm the conditions in Theorem 5. However, the function $f(z) = 0.5z^2$ does not meet the above conditions. In the same way, we can find that the function $f(z) = z^m/\{m(m-1)\}$ with a natural number $m \geq 2$ does not meet the conditions.*

We have a similar result for non-separable estimators. As for the estimator with γ-divergence, the loss function is convex if the equality $\delta = 0$ holds.

C ASYMPTOTICS UNDER MISSPECIFICATION

C.1 Misspecified Case

We have assumed that the model includes a true density. We can also consider a misspecified case, showing that the behavior of the proposed estimators associated with $f(x) = x\log x$, i.e., (11) is asymptotically the same as that of MLE. This implies that similar to the MLE, the proposed estimator with $f(x) = x\log x$ converges to the parameter that minimizes the KL-divergence between the model and the true distribution. Furthermore, its asymptotic variance is the same, even when the model is misspecified. We specifically have the following theorem.

Theorem 6 *Under certain regularity conditions, we have*

$$\sqrt{n}(\hat{\tau}_s - \tau^*) \xrightarrow{d} \mathcal{N}(0, \Omega_{1m}^{-1}\Omega_{2m}\Omega_{1m}^{-1}),$$

where $\tau^* = (c^*, \theta^*)$ is a value such that

$$\exp(c^*) = \int p(x;\theta^*)\mathrm{d}\mu(x),\ 0 = \mathrm{E}_*\{S(x;\theta)\},$$

and

$$\begin{aligned}\Omega_{1m} &= -\mathrm{E}_*\left[\left\{1 - \frac{q(x;\tau)}{\eta^*(x)}\right\}\nabla_{\tau^\top}\nabla_\tau \log q(x;\tau)|_{\tau^*}\right] \\ &\quad + \mathrm{E}_*\left\{\frac{q(x;\tau)}{\eta^*(x)}\nabla_\tau \log q(x;\tau)\nabla_{\tau^\top}\log q(x;\tau)|_{\tau^*}\right\}, \\ \Omega_{2m} &= \mathrm{Var}_*\left\{\nabla_\tau \log q(x;\tau)|_{\tau^*}\right\}.\end{aligned}$$

Two implications are observed in Theorem 6. First, this theorem is reduced to Theorem 2 when the model includes the true distribution, i.e., $\eta^*(x) = q(x;\tau^*)$. Second, the resulting form of Ω_{1m}, Ω_{2m} has a form similar to the terms that appear in the asymptotic result of MLE estimator when the model is normalized. For details, see Appendix C.1.

We have assumed that the model includes true density. In this section, we consider a misspecified case, showing that the behavior of the proposed estimators associated with KL divergence is asymptotically the same as that of MLE when $f(x) = x\log x$. This implies that similarly to MLE, the proposed estimator converges to the parameter that minimizes the KL-divergence between the model and the true distribution, even when the model is misspecified.

Before analyzing the proposed estimators, we review a misspecified case where the model can be normalized properly. The MLE under the misspecified model is equivalent to finding the closest model to the true distribution regarding KL divergence (White, 1982). The MLE estimator $\hat{\theta}_{\mathrm{MLE}}$ converges to the value maximizing the function $\theta \to \mathrm{E}_*\{\log p(x;\theta) - \log\int p(x;\theta)\mathrm{d}\mu(x)\}$. We denote this value as θ^*. The value θ^* satisfies the equation $\mathrm{E}_*\{S(x;\theta)\} = 0$, where $S(x;\theta)$ is

$$\nabla_\theta\left\{\log p(x;\theta) - \log\int p(x;\theta)\mathrm{d}\mu(x)\right\}.$$

It is well-known that the estimator $\hat{\theta}_{\mathrm{MLE}}$ has the following asymptotic property, that is, $\sqrt{n}(\hat{\theta}_{\mathrm{MLE}} - \theta^*)$ converges weakly to the normal distribution with mean 0 and variance

$$\mathrm{E}_*\{\nabla_{\theta^\top}S(x;\theta)|_{\theta^*}\}^{-1}\mathrm{Var}_*\{S(x;\theta)|_{\theta^*}\}\mathrm{E}_*\{\nabla_{\theta^\top}S(x;\theta)|_{\theta^*}\}^{-1}$$

The term $\mathrm{E}_*\{\nabla_{\theta^\top} S(x;\theta)|_{\theta^*}\}$ is

$$\mathrm{E}_*\left[\left\{1-\frac{\tilde{p}^*(x)}{\eta^*(x)}\right\}\nabla_{\theta^\top}\nabla_\theta \log p(x;\theta)|_{\theta^*}\right] + \mathrm{E}_*\left\{\frac{\tilde{p}^*(x)}{\eta^*(x)}\nabla_\theta \log p(x;\theta)\nabla_{\theta^\top}\log p(x;\theta)|_{\theta^*}\right\}$$
$$-\mathrm{E}_*\left\{\frac{\tilde{p}^*(x)}{\eta^*(x)}\nabla_\theta \log p(x;\theta)|_{\theta^*}\right\}\mathrm{E}_*\left\{\frac{\tilde{p}^*(x)}{\eta^*(x)}\nabla_{\theta^\top}\log p(x;\theta)|_{\theta^*}\right\},$$

where

$$\tilde{p}(x;\theta) = p(x;\theta)/\int p(x;\theta)\mathrm{d}\mu(x),$$

$\tilde{p}^*(x) = \tilde{p}(x;\theta^*)$. We also have

$$\mathrm{Var}_*\{S(x;\theta)|_{\theta^*}\} = \mathrm{Var}_*\{\nabla_\theta \log p(x;\theta)|_{\theta^*}\}.$$

Next, consider the asymptotic behavior of $\hat{\theta}_\mathrm{s}$ in (9) when the model is misspecified. We assume $f(x) = x\log x$, as in Example 3.1. In this case, the estimator $\hat{\theta}_\mathrm{s}$ converges in probability to θ^*, which satisfies the equation $\mathrm{E}_*[S(x;\theta)] = 0$. When $f(x)$ is not $x\log x$, a similar result can be obtained. However, the limits of estimators no longer converge to the same θ^*. With these settings, we have the following theorem.

Theorem 7 *Under regularity conditions as in Theorem 2, we have*

$$\sqrt{n}(\hat{\theta}_s - \theta^*) = \Omega_{1m}^{\dagger\,-1}\mathbb{G}_n\left\{\nabla_\theta \log p(x;\theta)|_{\tau^*}\right\} + \mathrm{o}_p(1),$$
$$\sqrt{n}(\hat{\theta}_s - \theta^*) \xrightarrow{d} \mathcal{N}(0, \Omega_{1m}^{\dagger\,-1}{\Omega^\dagger}_{2m}\Omega_{1m}^{\dagger\,-1}).$$

The specific forms of $\Omega_{1m}^\dagger$ *and* $\Omega_{2m}^\dagger$ *are*

$$\Omega_{1m}^\dagger = \mathrm{E}_*\left[\left\{1-\frac{q(x;\tau)}{\eta^*(x)}\right\}\nabla_{\theta^\top}\nabla_\theta \log p(x;\theta)|_{\tau^*}\right] + \mathrm{E}_*\left\{\frac{q(x;\tau)}{\eta^*(x)}\nabla_\theta \log p(x;\theta)\nabla_{\theta^\top}\log p(x;\theta)|_{\tau^*}\right\}$$
$$-\mathrm{E}_*\left\{\frac{q(x;\tau)}{\eta^*(x)}\nabla_\theta \log p(x;\theta)|_{\tau^*}\right\}\mathrm{E}_*\left\{\frac{q(x;\tau)}{\eta^*(x)}\nabla_{\theta^\top}\log p(x;\theta)|_{\tau^*}\right\},$$

and

$$\Omega_{2m}^\dagger = \mathrm{Var}\{\nabla_\theta \log p(x;\theta)|_{\theta^*}\}.$$

Two implications are observed in Theorem 7. First, when the model includes the true distribution, i.e., $\eta^*(x) = q(x;\tau^*)$, this theorem is reduced to Theorem 2. Second, the resulting form of ${\Omega^\dagger}_{1m}, {\Omega^\dagger}_{2m}$ has a form similar to terms appeared in the asymptotic result of the MLE estimator when the model is normalized.

C.2 Misspecified Poisson Model

Here, we examine the bahavior of each estimator when the model is misspecified. We assume unnormalized parametric models $P(x;\theta) \propto \exp(\theta)^x/x!, x \in \mathbb{N}_{\geq 0}$ based on Poisson distributions. We consider two scenarios based on the true distribution (well-specified case) $\exp(-2.0)2.0^x/x!$ and, (misspecified case) $0.5\exp(-2.0)2^{x-0.2}/(x-0.2)! + 0.5\exp(-1.0)/(x-1.2)!$.

We compared five estimators: **s-KL**, **s-Chi**, **s-JS**, **ns-**γ and **MLE**. The Monte Carlo mean and the standard error of KL divergence between true density and estimated density are presented in Table 5. This experiment reveals that the performance of each estimator significantly varies in the misspecified case, but not in the well-specified case. It is indicated that **s-KL** is preferable in terms of the KL divergence because it has a performance similar to that of MLE, even when the model is misspecified.

D Additional experiments

D.1 Performance of s-KL, s-Chi, s-JS in RBM

We compared four estimators: **s-JS**, **s-KL**, **s-Chi** and **MLE**. Refer to Table 6. It is shown that **s-KL** is generally stable.

Table 5: Monte Carlo mean and standard error of the KL divergence between the true density and estimated density scaled by sample size in a Poisson model. Parenthesis indicates a standard error.

well-specified case

n	s-KL	s-Chi	s-JS	ns-γ	MLE
1000	0.26	0.26	0.27	0.26	0.26
	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)
2000	0.25	0.26	0.25	0.25	0.25
	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)

misspecified case

n	s-KL	s-Chi	s-JS	ns-γ	MLE
1000	5.6	6.0	7.3	6.2	5.5
	(0.4)	(0.7)	(1.4)	(0.7)	(0.2)
2000	11.1	11.8	14.6	12.1	10.9
	(0.4)	(0.9)	(1.9)	(0.7)	(0.2)

Table 6: Monte Carlo mean and standard error of the KL divergence between the true density and estimated density scaled by sample size in RBM. Parenthesis indicates a standard error.

$\dim \boldsymbol{v} = 5, \dim \boldsymbol{h} = 2$, iteration: 50

n	s-JS	s-KL	s-Chi	MLE
100	5.91(2.66)	5.32(2.22)	6.73(4.07)	5.66(2.72)
500	4.94(2.02)	5.14(2.05)	6.88(3.03)	5.06(1.95)
1000	5.35(2.46)	5.43(2.58)	6.45(3.57)	5.57(2.74)

$\dim \boldsymbol{v} = 8, \dim \boldsymbol{h} = 2$, iteration: 20

n	s-JS	s-KL	s-Chi	MLE
500	26.3(12.9)	24.7(12.1)	30.2(12.3)	11.2(4.60)
1000	18.4(9.62)	14.6(9.28)	17.4(10.4)	8.38(3.09)
5000	10.5(3.73)	8.78(3.27)	18.9(7.35)	8.85(3.24)

Table 7: Median squared errors scaled by sample size

	n = 1000		n = 4000	
	Gaussian	Gamma	Gaussian	Gamma
MLE	0.24	14.3	0.26	14.8
NCE	0.26	54.4	0.28	61.2
s-KL	0.39	24.6	0.29	23.3
s-Chi	0.35	15.1	0.43	19.6
s-JS	0.48	16.3	0.26	18.5
ns-γ	0.75	14.5	0.35	36.5

D.2 Gaussian and Gamma Distributions

We perform toy experiments using Gaussian distribution and gamma distributions. These experiments show that proposed estimator's performance is almost the same as the MLE. In this section, we use median squared errors rather than mean squared errors.

Let us consider simple examples when the baseline measure is a Lebesgue measure. Here we define the following two

unnormalized models: Gaussian distribution, gamma distribution as follows:

$$p(x;\theta) = \exp(-\theta x^2),\ \widetilde{p}(x;\theta) = \sqrt{\frac{\theta}{\pi}}\exp(-\theta x^2),$$
$$p(x;\theta) = x^{\theta_1 - 1}\exp(-\theta_2 x),\ \widetilde{p}(x;\theta) = \frac{\theta_2^{\theta_1} x^{\theta_1-1}\exp(-\theta_2 x)}{\Gamma(\theta_1)}.$$

We write down each corresponding normalized model on the right side. Simulation is replicated for 100 times. Monte Carlo median squared errors are reported in Table D.2. Note that we use a half-normal distribution for the NCE in the case of the gamma distribution. It is indicated that proposed estimators have the similar performance as the MLE. This supports our theoretical result. However, it seems that each proposed estimator has a slightly different performance. One reason is that our analysis does not take high-order terms into account.

E PROOF OF THEOREMS

Proof of Theorem 1. Use Theorem 5.11 in van der Vaart (2002). We check two conditions; (1a) $\phi(x;\tau,\hat{\eta}_n)$ belongs to Glivenko Canteli class,(1b) for any $\epsilon > 0$, $\inf_{\{\tau:\|\tau-\tau^*)\|>\epsilon\}} \|\mathrm{E}_*\{\phi(x;\tau,\eta^*\})\|$. Regarding the first condition, we check in the proof of Theorem 2. Assumption (1b) is verified by the following two conditions: (1c) $\phi(x;\tau,\eta^*)$ is continuous with respect to τ, (1d) $\mathrm{E}_*\{\phi(x;\tau,\eta^*)\} = 0 \iff \tau = \tau^*$. The condition (1c) immediately holds assuming that $\theta \to p(x;\theta)$ is continuous. When the identification condition of the model $q(x;\tau_1) = q(x;\tau_2) \iff \tau_1 = \tau_2$ holds, (1d) is verified because $\mathrm{E}_*[\phi(x;\tau,\eta^*)] = 0 \iff q(x;\tau) = q(x;\tau^*) \iff \tau = \tau^*$ (Uehara et al., 2018). ■

Proof of Theorem 2.

First, under Assumptions 1-3 and (2a), we can check the following conditions;

- $(\tau,\eta) \to \phi(x;\tau,\eta)$ is continuous in an L^2 space $L^2(F_{\eta^*})$ at (τ^*,η^*)
- $\{\phi(x;\tau,\eta)\}$ belongs to a Donsker class
- $\hat{\tau}_{\mathrm{s}} \xrightarrow{p} \tau^*$
- Map $\tau \to \phi(x;\tau,\eta)$ is differentiable at τ^* uniformly in a neighrborhood of η^*
- The following matrix $\Omega = \mathrm{E}_*(\nabla_\tau \log q \nabla_{\tau^\top} \log q|_{\theta^*})$ is non-singular

to invoke Theorem 6.17. in van der Vaart (2002). Especially, the second condition is confirmed as follows. First, $\{q(x;\tau);\tau \in \Theta_\tau\}$ and $\{1(x \le t), t \in \mathbb{R}\}$ belong to Donsker class from Example 19.16 and Example 19.18 in (van der Vaart, 1998). Then, noting that

$$(q,\eta) \to \phi(q,\tau)$$

is a Lipsthicz continuous function, from Example 19.20 in (van der Vaart, 1998), $\phi(q(x);\eta(x)) = \phi(x;\tau,\eta)$ is also a Donsker class.

We have

$$\sqrt{n}(\hat{\tau}_{\mathrm{s}} - \tau^*) = \Omega^{-1}\mathbb{G}_n\left\{\nabla_\tau \log q(x;\tau)|_{\tau^*}\right\} + \mathrm{o}_p(1),$$
$$\sqrt{n}(\hat{\tau}_{\mathrm{s}} - \tau^*) \xrightarrow{d} \mathcal{N}(0,\Omega^{-1}).$$

The estimator $\hat{\tau}_{\mathrm{s}}$ is considered as the one satisfying $\mathbb{P}_n\phi(x;\tau,\eta)|_{\hat{\tau}_{\mathrm{s}},\hat{\eta}_n} = 0$, where $w(x) = q(x;\tau)/\eta(x)$ and $\phi(x;\tau,\eta)$ is

$$\left([f'\{h_1(w)\} - f'\{h_2(w)\}]\,h_1'(w) - f''\{h_2(w)\}h_2'(w)\,[h_1(w) - h_2(w)]\right) w\nabla_\tau \log q(x;\tau).$$

From Theorem 6.17. in van der Vaart (2002) based on the above assumptions, we have

$$\sqrt{n}(\hat{\tau}_{\mathrm{s}} - \tau^*) = -V^{-1}_{\tau^*,\eta^*}\sqrt{n}\mathrm{E}_*\{\phi(x)|_{\tau^*,\hat{\eta}_n}\} - V^{-1}_{\tau^*,\eta^*}\mathbb{G}_n\phi(x)|_{\tau^*,\eta^*} + \mathrm{o}_p(1 + \sqrt{n}\|\mathrm{E}_*[\phi(x)|_{\tau^*,\hat{\eta}_n}]\|), \tag{13}$$

where V_{τ^*,η^*} is a derivative of $\tau \to \mathrm{E}_*\{\phi(x;\tau,\eta^*)\}$ at τ^*. First, we calculate the derivative V_{τ^*,η^*}. The derivative is

$$\begin{aligned}
&\nabla_{\tau^\top}\mathrm{E}_*\{\phi(x;\tau,\eta^*)|_{\tau^*}\} = \mathrm{E}_*\{\nabla_{\tau^\top}\phi(x;\tau,\eta^*)|_{\tau^*}\}\\
&= \sqrt{n}\mathrm{E}_*[f''\{h_1(w)\}h_1'(w) - f''\{h_2(w)\}h_2'(w)]h_1'(w) - f''\{h_2(w)\}h_2'(w)\{h_1'(w) - h_2'(w)\}\\
&w\nabla_\tau \log q(x;\tau)\{\nabla_{\tau^\top} w\}\{\hat{\eta}_n(x) - \eta^*(x)\}|_{\tau^*,\eta^*}\\
&= f''(1)\{h_1'(1) - h_2'(1)\}^2\mathrm{E}_*(\nabla_\tau \log q \nabla_{\tau^\top}\log q|_{\tau^*})\\
&= f''(1)\{h_1'(1) - h_2'(1)\}^2\Omega.
\end{aligned}$$

Next consider each term in (13). The second term in (13) vanishes because $\phi(x)|_{\tau^*,\eta^*}$ is 0. Therefore, we only analyze the first term in (13):

$$\begin{aligned}
&\sqrt{n}\mathrm{E}_*\{\phi(x)|_{\tau^*,\hat{\eta}_n}\} = \sqrt{n}\mathrm{E}_*\{\phi(x)|_{\tau^*,\hat{\eta}_n}\} - \sqrt{n}\mathrm{E}_*\{\phi(x)|_{\tau^*,\eta^*}\}\\
&= \sqrt{n}\mathrm{E}_*[\nabla_\eta\phi(x)|_{\tau^*,\eta^*}\{\hat{\eta}_n(x) - \eta^*(x)\}] \qquad (14)\\
&+ \sqrt{n}\mathrm{E}_*\{\phi(x)|_{\tau^*,\hat{\eta}_n}\} - \sqrt{n}\mathrm{E}_*\{\phi(x)|_{\tau^*,\eta^*}\} - \sqrt{n}\mathrm{E}_*[\nabla_\eta\phi(x)|_{\tau^*,\eta^*}\{\hat{\eta}_n(x) - \eta^*(x)\}]. \qquad (15)
\end{aligned}$$

We decompose $\sqrt{n}\mathrm{E}_*\{\phi(x)|_{\tau^*,\hat{\eta}_n}\}$ into two terms again. The first term (14) is

$$\begin{aligned}
&\sqrt{n}\mathrm{E}_*[\nabla_\eta\phi(x)|_{\tau^*,\eta^*}\{\hat{\eta}_n(x) - \eta^*(x)\}]\\
=&\sqrt{n}\mathrm{E}_*[\{f''\{h_1(w)\}h_1'(w) - f''\{h_2(w)\}h_2'(w)\}h_1'(w) - f''\{h_2(w)\}h_2'(w)(h_1'(w) - h_2'(w)\}\{vgn\nabla_\eta w\}\\
&w\nabla_\tau \log q(x;\tau)|_{\tau^*,\eta^*}(\hat{\eta}_n(x) - \eta^*(x))]\\
=&-\sqrt{n}\int f''(1)\{h_1'(1) - h_2'(1)\}^2\frac{\nabla_\tau q(x;\tau)}{q(x;\tau)}|_{\tau^*}\{\hat{\eta}_n(x) - \eta^*(x)\}\mathrm{d}\mu(x)\\
=&-\sqrt{n}f''(1)\{h_1'(1) - h_2'(1)\}^2\mathbb{G}_n\left\{\frac{\nabla_\tau q(x;\tau)}{q(x;\tau)}|_{\tau^*}\right\}.
\end{aligned}$$

In addition, the second residual term (15) vanishes because, for some large C and $\tilde{\eta}$ is a between $\hat{\eta}_n$, we have

$$\begin{aligned}
&\|\sqrt{n}\mathrm{E}_*\{\phi_{\tau^*,\hat{\eta}_n}\} - \sqrt{n}\mathrm{E}_*\{\phi_{\tau^*,\eta^*}\} - \sqrt{n}\mathrm{E}_*[\nabla_\eta\phi(x;\tau,\eta)_{|\tau^*,\eta^*}\{\hat{\eta}_n(x) - \eta^*(x)\}]\|\\
&= \sqrt{n}\|\mathrm{E}_*[\nabla_{\eta\eta}\phi(x;\tau,\eta)_{|\tau^*,\tilde{\eta}}\{\hat{\eta}_n(x) - \eta^*(x)\}^2]\|\\
&\le C\sqrt{n}\|\mathrm{E}_*[\{\hat{\eta}_n(x) - \eta^*(x)\}^2].
\end{aligned}$$

From the second line to the third line, we use an assumption (2b). The last term goes to 0 in probability. By combining all things and substituting into (13), the statement is proved. ■

Proof of Corollary 1. The score function $S(x;\theta)$ can be written as

$$\nabla_\theta \log p(x;\theta) - \int \nabla_\theta \log p(x;\theta)\frac{p(x;\theta)}{\int p(x;\theta)\mathrm{d}\mu(x)}\mathrm{d}\mu(x).$$

Fisher information matrix $\mathfrak{I}_{\theta^*}^{-1}$ is $\mathrm{Var}_*\{S(x;\theta)|_{\theta^*}\}$, that is,

$$\mathrm{E}_*\{\nabla_\theta \log p(x;\theta)\nabla_{\theta^\top}\log p(x;\theta)|_{\theta^*}\} - \mathrm{E}_*\{\nabla_\theta \log p(x;\theta)|_{\theta^*}\}\mathrm{E}_*\{\nabla_{\theta^\top}\log p(x;\theta)|_{\theta^*}\}.$$

On the other hand, the component corresponding θ^* in Ω^{-1} can be also written as

$$\mathrm{E}_*\{\nabla_\theta \log p(x;\theta)\nabla_{\theta^\top}\log p(x;\theta)|_{\theta^*}\} - \mathrm{E}_*\{\nabla_\theta \log p(x;\theta)|_{\theta^*}\}\mathrm{E}_*\{\nabla_{\theta^\top}\log p(x;\theta)|_{\theta^*}\}$$

from Theorem 2 and Woodbury formula. This is the same as the Fisher information matrix. This concludes the proof. ■

Proof of Theorem 3.

We can do in the proof of Theorem 2. Here, $\eta(x)$ belongs to a Donsker class from Example 19.24 (van der Vaart, 1998); therefore, $\phi(x;\tau,\eta)$ belongs to Donsker class from Lipschitz continuous property. The problem is a drift term. We can derive the given theorem by calculating the drift term in the same way. The drift term $\sqrt{n}\mathrm{E}_*[\phi_{\tau^*,\hat{\eta}_n}]$ is decomposed into two terms, the main term:

$$\sqrt{n}\int \frac{\nabla_\tau q(x;\tau)}{q(x;\tau)}|_{\tau^*}\left\{\frac{1}{nh^{d_x}}\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)-\eta^*(x)\right\}\mathrm{d}\mu(x),$$

and the residual term. The main term corresponds to the term (14) in Theorem 2 and the residual term corresponds to the term (15) in Theorem 2. As revealed in the proof, the residual term is written as $\mathrm{O}_p(\sqrt{n}\|\hat{\eta}-\eta^*(x)\|^2)$. This term is equal to the order $\mathrm{o}_p(1)$ because $\frac{\nu}{2\nu+d_x} > 1/4$ holds from the assumption (2e). Next, we have

$$\sqrt{n}\int \frac{\nabla_\tau q(x;\tau)}{q(x;\tau)}|_{\tau^*}\left\{\frac{1}{nh^{d_x}}\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)\mathrm{d}\mu(x)-\mathrm{d}\mathbb{P}_n(x)\right\}=\mathrm{o}_p(1).$$

This holds from Theorem 8.11 in Newey and Mcfadden (1994) using assumptions (2d) and (2e). Then, the drift term becomes

$$\begin{aligned}\sqrt{n}\mathrm{E}_*[\phi_{\tau^*,\hat{\eta}_n}] &= \sqrt{n}\int \frac{\nabla_\tau q(x;\tau)}{q(x;\tau)}|_{\tau^*}\left\{\mathrm{d}\mathbb{P}_n(x)-\eta^*(x)\right\}\mathrm{d}\mu(x)+\mathrm{o}_p(1)\\ &= \mathbb{G}_n\left\{\frac{\nabla_\tau q(x;\tau)}{q(x;\tau)}|_{\tau^*}\right\}+\mathrm{o}_p(1).\end{aligned}$$

We have calculated the drift term. For the rest of the proof, it is the same as the proof in Theorem 2. ■

Proof of Theorem 4. We redefine $\sigma \equiv (\theta^\top, c_1, c_2)^\top$. To avoid abuse of notations, we write $U_{\alpha,\beta}(x;\sigma)$ as $U(x)$.

As in the proof of Theorem 2, we have

$$\sqrt{n}(\hat{\sigma}_{\text{ns-}\gamma}-\sigma^*) = -V^{-1}_{\sigma^*,\eta^*}\sqrt{n}\mathrm{E}_*\{U(x)|_{\sigma^*,\hat{\eta}_n}\}-V^{-1}_{\sigma^*,\eta^*}\mathbb{G}_nU(x)|_{\sigma^*,\eta^*}+\mathrm{o}_p(1+\sqrt{n}\|\mathrm{E}_*\{U(x)|_{\sigma^*,\hat{\eta}_n}\}\|), \tag{16}$$

where $\hat{\sigma}_{\text{ns-}\gamma}$ is a solution to $\tilde{\mathrm{E}}[U(x;\sigma)]=0$ and V_{σ^*,η^*} is a derivative of the map $\sigma \to \mathrm{E}_*\{U(x;\sigma,\eta^*)\}$ at σ^*.

First, we calculate the derivative V_{σ^*,η^*}. This becomes

$$\mathrm{E}_*\left[\begin{pmatrix}(\beta-\alpha)\nabla_\theta s(x;\theta)s(x;\theta)^\top & -s(x;\theta) & s(x;\theta)\\ (\beta-1)s(x;\theta)^\top\exp(c_1) & \exp(c_1) & 0\\ (\alpha-1)s(x;\theta)^\top\exp(c_2) & 0 & \exp(c_2).\end{pmatrix}\right],$$

which is evaluated at σ^* and $s(x;\theta)=\nabla_\theta\log p(x;\theta)$. The term corresponding θ in the above matrix $V^{-1}_{\sigma^*,\eta^*}$ is

$$\begin{aligned}&(\beta-\alpha)^{-1}\left[\mathrm{E}_*\{\nabla_\theta\log p(x;\theta)\nabla_{\theta^\top}\log p(x;\theta)\}-\mathrm{E}_*\{\nabla_\theta\log p(x;\theta)\}\mathrm{E}_*\{\nabla_{\theta^\top}\log p(x;\theta)\}\right]^{-1}|_{\theta^*}\\ &=(\beta-\alpha)^{-1}\mathfrak{J}^{-1}_{\theta^*}.\end{aligned}$$

Then, we analyze each term in (16). First of all, the second term in (16) becomes zero because $U(x;\sigma^*,\eta^*)=0$. Therefore, we only consider the first term in (16). We have

$$\begin{aligned}\sqrt{n}\mathrm{E}_*\{U(x)|_{\sigma^*,\hat{\eta}_n}\} &= \sqrt{n}\mathrm{E}_*[\nabla_\eta U(x)_{|\sigma^*,\eta^*}\{\hat{\eta}_n(x)-\eta^*(x)\}]+\mathrm{o}_p(1)\\ &= \sqrt{n}\mathrm{E}_*\left[\begin{pmatrix}(\alpha-\beta)\frac{\nabla_\theta\log p(x;\theta)}{\eta^*(x)}\\ (\beta-1)/\eta^*(x)\\ (\alpha-1)/\eta^*(x)\end{pmatrix}\{\hat{\eta}_n(x)-\eta^*(x)\}\right]+\mathrm{o}_p(1).\end{aligned}$$

Therefore, the first term corresponding θ in the above equation becomes

$$\begin{aligned}&\mathfrak{J}_{\theta^*}^{-1}\sqrt{n}\int \nabla_\theta \log p(x;\theta)|_{\theta^*}\{\hat{\eta}_n(x)-\eta^*(x)\}\mathrm{d}\mu(x)\\&=\mathfrak{J}_{\theta^*}^{-1}\mathbb{G}_n\left\{\nabla_\theta \log p(x;\theta)|_{\theta^*}\right\}.\end{aligned}$$

Finally, we get

$$\sqrt{n}(\hat{\theta}_{\text{ns-}\gamma}-\theta^*)=\mathfrak{J}_{\theta^*}^{-1}\mathbb{G}_n\{\nabla_\theta \log p(x;\theta)|_{\theta^*}\}+\mathrm{o}_p(1).$$

■

Proof of Theorem 5. Let us define $\ell_i(\tau)$ as the loss for the sample x_i, i.e.,

$$\ell_i(\tau)=-f'(z_i)+w_i f'(z_i)-f(z_i),$$

where $z_i=q(x_i;\tau)/\hat{\eta}(x_i)$. The loss function is expressed by the total sum of $\ell_i(\tau)$ over all samples. For the unnormalized exponential model, some calculation yields the Hessian matrix of $\ell_i(\tau)$,

$$\begin{aligned}\nabla^2\ell_i(\tau)&=\left[f''(z_i)z_i^2+(z_i-1)\left\{f'''(z_i)z_i^2+f''(z_i)z_i\right\}\right]\phi(x_i)\phi(x_i)^\top\\&=z_i\left\{(2z_i-1)f''(z_i)+z_i(z_i-1)f'''(z_i)\right\}\phi(x_i)\phi(x_i)^\top.\end{aligned}$$

The assumption of the theorem guarantees that the coefficient above is non-negative; hence, the Hessian matrix of $\ell_i(\tau)$ is non-negative definite, so is the loss function. Eventually, the loss function is convex in the parameter τ. ■

Proof of Theorem 6. The estimator $\hat{\tau}_{\mathrm{s}}$ can be considered as the one satisfying $\mathbb{P}_n\phi_{\hat{\tau}_{\mathrm{s}},\hat{\eta}_n}=0$, where

$$\phi(x;\tau,\eta)=\nabla_\tau \log q(x;\tau)-\left\{\frac{q(x;\tau)}{\eta(x)}\right\}\nabla_\tau \log q(x;\tau).$$

From Theorem 6.17. van der Vaart (2002), we have

$$\sqrt{n}(\hat{\tau}_{\mathrm{s}}-\tau^*)=-V_{\tau^*,\eta^*}^{-1}\sqrt{n}\mathrm{E}_*\{\phi|_{\tau^*,\hat{\eta}_n}\}-V_{\tau^*,\eta^*}^{-1}\mathbb{G}_n\phi(x_i;\tau^*,\eta^*)+\mathrm{o}_p(1+\sqrt{n}\|\mathrm{E}_*[\phi|_{\tau^*,\hat{\eta}_n}]\|),\tag{17}$$

where V_{τ^*,η^*} is a derivative of $\tau\to\mathrm{E}_*\{\phi(x;\tau,\eta^*)\}$ at η^*.

First, we will see a more specific form of τ^*. The value τ^* satisfy the equation $\mathrm{E}_*\{\phi(x;\tau,\eta^*)\}=0$. Noting that $\nabla_{\tau^\top}\log q(x;\tau)=(1,\nabla_{\theta^\top}\log p(x;\theta))$, we can get the form of c^* and θ^* specified in the statement.

Next, we calculate the derivative V_{τ^*,η^*}. The derivative is

$$\begin{aligned}&\nabla_{\tau^\top}\mathrm{E}_*\{\phi(x;\tau,\eta)|_{\tau^*}\}\\&=\mathrm{E}_*\{\nabla_{\tau^\top}\phi(z;\tau,\eta)|_{\tau^*}\}\\&=\mathrm{E}_*\left[\left\{-1+\frac{q(x;\tau)}{\eta(x)}\right\}\nabla_{\tau^\top}\nabla_\tau \log q(x;\tau)|_{\tau^*}\right]-\mathrm{E}_*\left\{\frac{q(x;\tau)}{\eta(x)}\nabla_\tau \log q(x;\tau)\nabla_{\tau^\top}\log q(x;\tau)\right\}|_{\tau^*}\\&=-\Omega_1.\end{aligned}$$

Next, consider each term in (17). The second term is

$$\Omega_1^{-1}\mathbb{G}_n\left[\left\{1-\frac{q(x;\tau)}{\eta(x)}\right\}|_{\tau^*,\eta^*}\nabla_\tau \log q(x;\tau)|_{\tau^*}\right].$$

The first term is

$$\begin{aligned}
\sqrt{n}\Omega_{1m}^{-1}\mathrm{E}_*(\phi|_{\tau^*,\hat{\eta}_n}) &= \sqrt{n}\Omega_{1m}^{-1}\mathrm{E}_*(\phi|_{\tau^*,\hat{\eta}_n}) - \sqrt{n}\mathrm{E}_*(\phi|_{\tau^*,\eta^*}) \\
&= \sqrt{n}\Omega_{1m}^{-1}\mathrm{E}_*\left[\nabla_\eta \phi(x)_{|\tau^*,\eta^*}\{\hat{\eta}_n(x) - \eta^*(x)\}\right] + \mathrm{o}_p(1) \\
&= \sqrt{n}\Omega_{1m}^{-1}\mathrm{E}_*\left[\frac{q(x;\tau)}{\eta^2(x)}\nabla_\tau \log q(x;\tau)|_{\tau^*,\eta^*}\{\hat{\eta}_n(x) - \eta^*(x)\}\right] + \mathrm{o}_p(1) \\
&= \sqrt{n}\Omega_{1m}^{-1}\int \frac{q(x;\tau)}{\eta(x)}|_{\tau^*,\eta^*}\nabla_\tau \log q(x;\tau)\{\hat{\eta}_n(x) - \eta^*(x)\}\mathrm{d}\mu(x) + \mathrm{o}_p(1) \\
&= \Omega_{1m}^{-1}\mathbb{G}_n\left\{\frac{q(x;\tau)}{\eta(x)}|_{\tau^*,\eta^*}\nabla_\tau \log q(x;\tau)|_{\tau^*}\right\} + \mathrm{o}_p(1).
\end{aligned}$$

Adding the first term and the second term, we get

$$\sqrt{n}(\hat{\tau}_{\mathrm{s}} - \theta^*) = \Omega_{1m}^{-1}\mathbb{G}_n\left\{\nabla_\tau \log q(x;\tau)|_{\tau^*,\eta^*}\right\} + \mathrm{o}_p(1).$$

Therefore, we conclude that $\sqrt{n}(\hat{\tau}_{\mathrm{s}} - \theta^*)$ converges to the normal distribution $\mathcal{N}(0, \Omega_{1m}^{-1}\Omega_{2m}\Omega_{1m}^{-1})$. ■

Proof of Theorem 7.

We calculate matrix, corresponding to θ term in Theorem 6. The matrix Ω_{1m} in Theorem 6 is equal to the following block matrix:

$$\begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix},$$

where $\Omega_{11} = 1$,

$$\Omega_{21} = \mathrm{E}_*\left\{\nabla_\tau \log q(x;\tau)\frac{q(x;\tau)}{\eta^*(x)}|_{\tau^*}\right\}$$

and

$$\Omega_{22} = \mathrm{E}_*\left[\left\{1 - \frac{q(x;\tau)}{\eta^*(x)}\right\}\nabla_{\theta^\top}\nabla_\theta \log q(x;\tau)|_{\tau^*}\right] + \mathrm{E}_*\left\{\frac{q(x;\tau)}{\eta^*(x)}\nabla_\theta \log q(x;\tau)\nabla_{\theta^\top}\log q(x;\tau)|_{\tau^*}\right\}.$$

From Woodbury formula, the corresponding term to θ in Ω_{1m}^{-1} is $\Omega_{1m}^\dagger$ where

$$\begin{aligned}
\Omega_{1m}^\dagger = &\mathrm{E}_*\left[\left\{1 - \frac{q(x;\tau)}{\eta^*(x)}\right\}\nabla_{\theta^\top}\nabla_\theta \log p(x;\theta)|_{\tau^*}\right] + \mathrm{E}_*\left\{\frac{q(x;\tau)}{\eta^*(x)}\nabla_\theta \log p(x;\theta)\nabla_{\theta^\top}\log p(x;\theta)|_{\tau^*}\right\} \\
&- \mathrm{E}_*\left\{\frac{q(x;\tau)}{\eta^*(x)}\nabla_\theta \log p(x;\theta)|_{\tau^*}\right\}\mathrm{E}_*\left\{\frac{p(x;\theta)}{\eta^*(x)}\nabla_{\theta^\top}\log p(x;\theta)|_{\tau^*}\right\}.
\end{aligned}$$

On the other hand, the corresponding part in Ω_{2m} is $\Omega_{2m}^\dagger$, where

$$\Omega_{2m}^\dagger = \mathrm{Var}\{\nabla_\theta \log p(x;\theta)|_{\theta^*}\},$$

noting that $\nabla_{\tau^\top}\log q(x;\tau) = \{1, \nabla_{\theta^\top}\log p(x;\theta)\}$. This concludes the proof. ■

Note the difference between the normalized case and unnormalized case is that $\tilde{p}(x;\theta^*)/\eta^*$ is used when the model is normalized; while, $q(x;\tau^*)/\eta^*$ is used in $\Omega_{1m}^\dagger$ and $\Omega_{2m}^\dagger$ when the model is unnormalized.