Yale University, Department of Statistics

Variational inference via score matching

Natalie Doss

August 15, 2019

Contents

1 Introduction

Score matching is a useful technique in density estimation. Let $p(z)$ be the true data density. Score matching seeks a density q to minimize a Fisher divergence, defined via

$$\mathcal{L}(q) = \mathbb{E}_p \|\nabla_z \log q(z) - \nabla_z \log p(z)\|_2^2. \tag{1}$$

This objective does not require estimation of the normalizing constant, since we take $\nabla_z \log q(z)$. By a simple argument using integration by parts, this objective simplifies to something that we can easily optimize in q; see Lemma 3.1. For exponential families, it further simplifies into a nice quadratic; see Lemma 3.2. It has been used in high dimensional density estimation where the normalizing constant is not computable, for example, in graphical models, c.f. Hyvarinen (2005), Lin *et al.* (2016), and Janofsky (2015).

We notice that the objective (1) is usable for posterior inference. Suppose we observe data x generated according to the conditional density $p(x|z)$. Associated to x is a latent variable z with a prior $p(z)$. We seek the posterior density:

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}.$$

But typically $p(x)$ is not tractable. But minimizing the following objective in q would not require us to have $p(x)$, since $\nabla_z \log p(z|x) = \nabla_z \log p(x|z) + \nabla_z \log p(z)$.

$$\mathcal{L}(q) = \mathbb{E}_{p(z|x)} \|\nabla_z \log q(z|x) - \nabla_z \log p(z|x)\|_2^2.$$

But is this something we can minimize? It turns out that it is. Suppose we assume that $p(z|x) \propto e^{f(z|x)}$, i.e., the posterior has an exponential family form. Suppose further that $f(z|x) = f_x(z) = \gamma^\top \phi(z)$, where $\gamma \in \mathbb{R}^K$ and $\phi : \mathbb{R}^d \to \mathbb{R}^K$. Then just as in Lemma 3.2, the objective simplifies to:

$$\int_z p(z|x) \left(\frac{1}{2} \gamma' A(z) \gamma + \gamma' k(z) \right) \, dz \tag{2}$$

which has optimal solution

$$\hat{\gamma} = \bar{A}^{-1} \bar{k} \tag{3}$$

where

$$\bar{A} = \int p(z, x) A(z) \, dz$$
$$\bar{k} = \int p(z, x) k(z) \, dz$$

where $\bar{A}, \bar{k}$ are defined below. Note the following key fact: the optimal solution requires integrals only over $p(z, x)$, not over $p(z|x)$, since the form in (3) allows the normalizing constant $p(x)$ to drop out! Thus this is something we can optimize.

But here are some important caveats. It actually is difficult to compute these integrals over the joint density. If we could, computing $p(x)$ wouldn't be a problem in the first place. We have considered importance-sampling. One issue is that often $p(x|z)$ is extremely small if n is large. Another is that, e.g., in the Gaussian family case, we still need $p(x)$; see Example 1. In fact, in any case where the first or second derivatives are constant, again, we would just be computing $p(x)$ via importance sampling.

There are other alternatives for how to optimize a score-matching type of divergence. Ranganath *et al.* (2016) use the objective:

$$\mathcal{L}(q) = \mathbb{E}_q \|\nabla_z \log q(z) - \nabla_z \log p(z)\|_2^2. \tag{4}$$

They re-express it in operator variational inference form, i.e., in the form $\mathbb{E}_q O(f, p)$. They use the Langevin Stein operator and iterate over optimizing the function f and the density q; see Section 5 for details. This has the weakness that it requires a min-max type of optimization.

An equivalence between the objectives (4) and (1) is in Lemma 5.3. Instead of the approach in Ranganath *et al.* (2016), we might directly optimize (4). We carry this out in Tensorflow.

2 Examples

Example 1 (Score-matching with $\mathbb{E}_p$ for posterior inference when q has Gaussian form)**.** *Suppose q has form $N(\mu_x, \Sigma_x)$ where $\mu_x \in \mathbb{R}^d, \Sigma_x \in \mathbb{R}^{d\times d}$. Let our prior $p(z)$ be the $N(0, I_d)$ density. Note that we could have $d \gg n$. We have $q(z) \propto e^{g(z)}$ where*

$$g(z) = \gamma' \phi(z)$$

where $K = 2d$ and

$$\gamma = \left(\Sigma^{-1}, \Sigma^{-1}\mu\right)^T$$
$$\phi(z) = \left(-\frac{1}{2}zz', z\right)^T$$

So (for $d = 1$, to keep it simple):

$$\frac{\partial \phi(z)}{\partial z} = (-z, 1)^T$$
$$\frac{\partial^2 \phi(z)}{\partial z^2} = (1, 0)^T$$

We have $A(z) \in \mathbb{R}^{2d\times 2d}$. Let e.g. z^2 indicate $(z_1^2, \dots, z_d^2)$. We have

$$A(z) = \begin{pmatrix} diag(z^2) & diag(-z) \\ diag(-z) & diag(1) \end{pmatrix} \text{ and}$$
$$k(z) = (rep(-1, d), rep(0, d))^T$$

Each submatrix is in $\mathbb{R}^{d\times d}$. We parameterize γ via B; it is some multi-layer non-linear function with many parameters B. Our objective, written as a sum, is

$$\mathbb{E}_{p(z|x)}\left(\sum_{j\le d}\gamma_{1j}^2 z_j^2 - \sum_{j\le d}\gamma_{2j}\gamma_{1j}z_j + \sum_{j\le d}\gamma_{2j}^2 - \sum_{j\le d}\gamma_{1j}\right)$$

Note that as noted in **?** *and Lin et al. (2016), there are closed-form solutions of μ, Σ. To see it, note that if $q(z)$ is $N(\mu, \Sigma)$,*

$$\log q(z) \propto \frac{-(z-\mu)'\Sigma^{-1}(z-\mu)}{2}$$

So the score-matching objective is

$$\frac{1}{2}\|\nabla_z \log q(z)\|_2^2 + \Delta_z \log q(z) = \frac{1}{2}\|\Sigma^{-1}(z-\mu)\|_2^2 + tr(\Sigma^{-1})$$

We can directly obtain:

$$\hat{\mu} = \mathbb{E}_{p(z|x)} z$$
$$\hat{\Sigma}_i = \mathbb{E}_{p(z|x)}(z - \hat{\mu})(z - \hat{\mu})'$$

While this is estimable for density estimation, it requires the posterior for us. Why does this happen when we can still not require $p(x)$ if we compute the exponential family parameter? Notice that e.g. in the simple $d = 1$ case, the quadratic we'd obtain for score-matching is:

$$\gamma' \mathbb{E}_{p(z|x)} \begin{pmatrix} z^2 & z \\ z & 1 \end{pmatrix} \gamma - \gamma' \mathbb{E}_{p(z|x)} \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Optimizing this allows us to remove the $p(x)$, but notice that then, we optimize in γ $\gamma' \bar{A} \gamma' - \gamma' \bar{k}$, the usual thing, where

$$\bar{k} = \begin{pmatrix} p(x) \\ 0 \end{pmatrix}$$

Now again we can approximate $p(x)$ via importance sampling, but we do have to approximate it either way.

3 Basic score matching

Lemma 3.1 (Score-matching simplifies to something simple using integration by parts)**.** *Let $\hat{q} = \operatorname{argmin}_q \mathcal{L}(q)$ where $\mathcal{L}(q)$ is as above. Then*

$$\begin{aligned} \hat{q} &= \operatorname*{argmin}_q \int_z p(z|x) \sum_{i \le d} \frac{1}{2} \left(\frac{\partial \log q(z)}{\partial z_i} \right)^2 + \frac{\partial^2 \log q(z)}{\partial z_i^2} \; dz \\ &= \operatorname*{argmin}_q \int_z p(z|x) \left(\frac{1}{2} \|\nabla \log q\|_2^2 + \triangle \log q \right) \; dz \end{aligned}$$

Proof. I just write this for one dimension. For an extension to higher dimensions, see Hyvarinen (2005). I sometimes write $p'(z)$ instead of $\frac{\partial p(z)}{\partial z}$.

$$\int p(z) \left(\frac{\partial \log p(z)}{\partial z} - \frac{\partial \log q(z)}{\partial z} \right)^2 \; dz \overset{c}{=} \int p(z) \left(\frac{\partial \log q(z)}{\partial z} \right)^2 - 2 \int p'(z) \frac{\partial \log q(z)}{\partial z}$$

And using integration by parts,

$$\int p'(z) \frac{\partial \log q(z)}{\partial z} \overset{c}{=} - \int p(z) \frac{\partial^2 q(z)}{\partial z^2}$$

□

Lemma 3.2 (The score-matching objective simplifies into a nice quadratic for exponential families.)**.** *Let*

$$q(z) = \frac{e^{g(z)}}{\int e^{g(z)} dz}$$

Suppose $g : \mathbb{R}^d \to \mathbb{R}$, with

$$g(z) = \gamma' \phi(z)$$

where $\phi : \mathbb{R}^d \to \mathbb{R}^K$ and $\gamma \in \mathbb{R}^K$. Write $\phi(z) = (\phi_1(z), \ldots, \phi_K(z))'$. Estimating q means estimating γ. Define

$$\frac{\partial \phi(z)}{\partial z_i} = \left(\frac{\partial \phi_1(z)}{\partial z_i}, \ldots, \frac{\partial \phi_K(z)}{\partial z_i} \right)^T$$
$$\frac{\partial^2 \phi(z)}{\partial z_i^2} = \left(\frac{\partial^2 \phi_1(z)}{\partial z_i^2}, \ldots, \frac{\partial^2 \phi_K(z)}{\partial z_i^2} \right)^T$$

For a non-bounded density (the second line is for an exponential family), the objective is, as in Janofsky (2015) and Hyvarinen (2007):

$$\begin{aligned} h_\gamma(z) &= \sum_{i \le d} \frac{1}{2} \left(\frac{\partial g(z)}{\partial z_i} \right)^2 + \frac{\partial^2 g(z)}{\partial z_i^2} \\ &= \sum_{i \le d} \frac{1}{2} \left(\gamma' \frac{\partial \phi(z)}{\partial z_i} \right)^2 + \gamma' \frac{\partial^2 \phi(z)}{\partial z_i^2} \end{aligned}$$

Now here are the definitions of A, k. Note $A \in \mathbb{R}^{K \times K}, k \in \mathbb{R}^K$. For a non-bounded density:

$$\begin{aligned} k(z) &= \sum_{i \le d} \frac{\partial^2 \phi(z)}{\partial z_i^2} \\ A(z) &= \sum_{i \le d} \frac{\partial \phi(z)}{\partial z_i} \frac{\partial \phi(z)}{\partial z_i}' \end{aligned}$$

3.1 Bounded density

For a bounded density, score matching has a slightly different objective. See Janofsky (2015) for more details.

$$\begin{aligned} k_1(z) &= \sum_{i \le d} 2(2z_i - 1) z_i (1 - z_i) \frac{\partial \phi(z)}{\partial z_i} \\ k_2(z) &= \sum_{i \le d} z_i^2 (1 - z_i)^2 \frac{\partial^2 \phi(z)}{\partial z_i^2} \\ k(z) &= k_1(z) - k_2(z) \\ A(z) &= \sum_{i \le d} \frac{\partial \phi(z)}{\partial z_i} \frac{\partial \phi(z)}{\partial z_i}' z_i^2 (1 - z_i)^2 \end{aligned}$$

For a bounded density (the second line is for an exponential family), the objective is:

$$
\begin{aligned}
h_\gamma(z) &= \sum_{i\le d} \frac{1}{2}\left(\frac{\partial g(z)}{\partial z_i} z_i(1-z_i)\right)^2 - 2(2z_i-1)z_i(1-z_i)\frac{\partial g(z)}{\partial z_i} + z_i^2(1-z_i)^2\frac{\partial^2 g(z)}{\partial z_i^2} \\
&= \sum_{i\le d}\left(\gamma'\frac{\partial \phi(z)}{\partial z_i} z_i(1-z_i)\right)^2 - 2(2z_i-1)z_i(1-z_i)\gamma'\frac{\partial \phi(z)}{\partial z_i} + z_i^2(1-z_i)^2\gamma'\frac{\partial^2 \phi(z)}{\partial z_i^2}
\end{aligned}
$$

In one dimension, and for $z \in [0,1]$, this simplifies to the following. Let $\phi : [0,1] \to \mathbb{R}^K$.

$$
\begin{aligned}
h_\gamma(z) &= \frac{1}{2}\left(\frac{\partial g(z)}{\partial z} z(1-z)\right)^2 - 2(2z-1)z(1-z)\frac{\partial g(z)}{\partial z} + z(1-z)\frac{\partial^2 g(z)}{\partial z^2} \\
&= \frac{1}{2} z^2(1-z)^2\gamma' A(z)\gamma - 2(2z-1)z(1-z)\gamma' k_1(z) + z^2(1-z)^2\gamma' k_2(z)
\end{aligned}
$$

where

$$
\begin{aligned}
k_1(z) &= \left(\frac{\partial \phi_1(z)}{\partial z}, \dots, \frac{\partial \phi_K(z)}{\partial z}\right)' \\
k_2(z) &= \left(\frac{\partial^2 \phi_1(z)}{\partial z^2}, \dots, \frac{\partial^2 \phi_K(z)}{\partial z^2}\right)' \\
k(z) &= k_1(z) - k_2(z) \\
A(z) &= k_1(z) k_1(z)'
\end{aligned}
$$

4 Score-matching with $\mathbb{E}_q$

The following lemma shows that the score-matching objective (4) decomposes into two components: a component matching q to p, and a component restricting q to be reasonably smooth. Thus the score-matching divergence decomposes in a similar manner as the KL divergence.

Lemma 4.1. *Suppose $q(z) = \exp(f(z) - \Psi)$ where Ψ is the normalizing constant. Let $\mathcal{L}(q) = \mathbb{E}_q\|\nabla_z \log p - \nabla_z \log q\|_2^2$. Then*

$$
\mathcal{L}(q) =^c \frac{1}{2}\mathbb{E}_q\|\nabla_z \log p(z,x)\|_2^2 + \mathbb{E}_q \Delta \log p(z,x) + \mathbb{E}_q\|\nabla_z f(z)\|_2^2
$$

Proof. [Only in one dimension for now.]

$$
\mathcal{L}(q) = \int q(z)\left(\frac{\partial}{\partial z}\log p\right)^2 - 2\int q(z)\frac{q'(z)}{q(z)}\frac{p'(z)}{p(z)} + \int q(z)\left(\frac{\partial}{\partial z}\log q(z)\right)^2
$$

The third term is $\mathbb{E}_q f'(z)^2$. And for the middle term, using integration by parts,

$$
\int q(z)\frac{q'(z)}{q(z)}\frac{p'(z)}{p(z)} = \int q'(z)\frac{\partial}{\partial z}\log p(z) =^c -\int q(z)\frac{\partial^2}{\partial z^2}\log p(z)
$$

□

The following lemma shows that when we assume $q(z)$ is a Gaussian density, the score-matching objective as in (4) reduces to the Laplace method as in Wang & Blei (2013). In retrospect, this is obvious by Lemma 5.1; since we assume a Gaussian form for the model, as long as we are optimizing over $\mathcal{Q}$ that includes Gaussian densities, we should obtain "the right thing." But this is nice as a sanity check.

Lemma 4.2. *Let $z \in \mathbb{R}^d$. Let $f(z) := \log p(z)$ and suppose for any z_0,*

$$f(z) \approx f(z_0) + \nabla f(z_0)^T (z - z_0) + \frac{1}{2}(z - z_0)^T \nabla^2 f(z_0)(z - z_0)$$

. Let $z^* \in \operatorname{argmax} f(z)$. Let $\mathcal{Q}$ be $N(\lambda, V)$ densisties. Let $\hat{q} = \operatorname{argmin}_{q \in \mathcal{Q}} \mathbb{E}_q \|\nabla_z \log p - \nabla_z \log q\|_2^2$. Then $\hat{q}$ is the $N(\hat{\lambda}, \hat{V})$ density with

$$\begin{aligned}\hat{\lambda} &= z^* \\ \hat{V} &= -\nabla^2 f(z^*)\end{aligned}$$

Proof. Let z^* be a mode of f, i.e., $\nabla_z f(z^*) = 0$. Then

$$f(z) \approx f(z_0) + \frac{1}{2}(z - z^*)^T \nabla^2 f(z^*)(z - z^*)$$

So

$$\nabla_z f(z) \approx \nabla^2 f(z^*)(z - z^*)$$

Let q be the $N(\lambda, V)$ density where $\lambda \in \mathbb{R}^d$ and $V \in \mathbb{R}^{d \times d}$; λ and V are the parameters to be optimized. Now $\nabla_z \log q(z) = -V^{-1}(z - \lambda)$. Now $z = V^{1/2} z_0 + \lambda$, where $z_0 \sim N(0, I_p)$. So

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_q \|\nabla_z \log q(z) - \nabla_z \log p(z)\|_2^2 \\ &= \mathbb{E}_{z_0 \sim N(0, I_p)} \|-V^{-1/2} z_0 - \nabla^2 f(z^*)(V^{1/2} z_0 + \lambda - z^*)\|_2^2 \\ &= \mathbb{E}_{z_0} \left\| \left(-V^{-1/2} - \nabla^2 f(z^*) V^{1/2}\right) z_0 - \nabla^2 f(z^*)(\lambda - z^*)\right\|_2^2 \\ &= \mathbb{E}_{z_0} \left\| \left(-V^{-1/2} - \nabla^2 f(z^*) V^{1/2}\right) z_0 \right\|_2^2 + \|\nabla^2 f(z^*)(\lambda - z^*)\|_2^2 \\ &= \|V^{-1/2} + \nabla^2 f(z^*) V^{1/2}\|_F^2 + \|\nabla^2 f(z^*)(\lambda - z^*)\|_2^2\end{aligned}$$

Optimizing in λ yields:

$$\hat{\lambda} = z^*$$

Let $M = \nabla^2 f(z^*)$. Let's do one dimension to check:

$$v^{-1/2} + m v^{1/2} = \frac{1 + mv}{\sqrt{v}}$$

And

$$\left(\frac{1 + mv}{\sqrt{v}}\right)^2 = \frac{1 + 2mv + m^2 v^2}{v} = \frac{1}{v} + 2m + m^2 v$$

Now

$$\frac{\partial}{\partial v}\left(\frac{1}{v}+2m+m^2v\right)=-\frac{1}{v^2}+m^2$$

Setting this equal to zero yields:

$$V^2=M^{-2}\Rightarrow V=\pm M^{-1}$$

which is exactly the Laplace method solution. □

Example 2 (Lemma 4.1 in the Gaussian case)**.** *Consider the objective (4). In the case where q is Gaussian, as in, Lemma 4.2, we can see that only optimizing the first two terms of the score-matching objective would lead to a terribly sharp q (with zero variance). The final term (the penalty on the smoothness of q) forces more smoothness. Let the setting be as in Lemma 4.2: q is the $N(\lambda, V)$ and the Taylor expansion for $\log p(x,z)$ holds. Now,*

$$\begin{aligned}\frac{1}{2}\mathbb{E}_q\|\nabla_z \log p(z,x)\|_2^2+\mathbb{E}_q\Delta\log p(z,x) &= \frac{1}{2}\mathbb{E}_q(z-z^*)^T\left(\nabla^2 f(z^*)\right)^2(z-z^*)-\mathbb{E}_q\nabla^2 f(z^*)\\ &=^c \frac{1}{2}tr\left(V^T\left(\nabla^2 f(z^*)\right)^2 V\right)+(\lambda-z^*)^T\left(\nabla^2 f(z^*)\right)^2(\lambda-z^*)\end{aligned}$$

which would be optimized by setting $\hat{\lambda}=z^*$ and $\hat{V}=0$. But we have the penalty:

$$\mathbb{E}_q\|\nabla\log q\|_2^2=tr(V^{-1})$$

Due to this penalty, we end up with a reasonable q.

5 Relationship between score-matching, Stein discrepancy, and OVI

Let $\psi_p(z)=\frac{\partial \log p(z)}{\partial z}=p'(z)/p(z)$ and similarly for $\psi_q(z)$. Let $\triangle p=\frac{\partial^2\log p(z)}{\partial z^2}$. Let $O^{p,q}$ be some operator depending on p,q. I will use the following form for exponential families in z:

$$p(z)=\exp\left(\gamma\phi(z)-\Psi(\gamma)\right) \tag{5}$$

The authors of Ranganath *et al.* (2016) propose to optimimze following variational objective in q:

$$\sup_{f\in\mathcal{F}}|\mathbb{E}_{q(z)}(O^p_{LS}f)(z)| \tag{6}$$

where

$$O^p_{LS}f=\psi_p f+\nabla f \tag{7}$$

Just as in the proof of Lemma 5.2, the objective (6) is

$$\begin{aligned}\sup_{f\in\mathcal{F}}|\mathbb{E}_q O^p_{LS}f| &= \sup_{f\in\mathcal{F}}\mathbb{E}_{q(z)}(\psi_q(z)-\psi_p(z))f(z) && (8)\\ &=\mathbb{E}_{q(z)}(\psi_q(z)-\psi_p(z))^2 && (9)\end{aligned}$$

i.e., their objective is the score-matching objective with expectation over q. But if we use a specific class $\mathcal{F}$ as in the following lemma, this becomes the usual score-matching objective with expectation over p.

The following lemma shows that (6) is valid in that it is zero when $q = p$. Of course, when we express the objective as in Lemma 5.2, this becomes even more obvious.

Lemma 5.1. *Assume q is zero at the boundaries of the sample space. Then*

$$\mathbb{E}_{q(z)}\left(\psi_q f(z) + f'(z)\right) = 0$$

Proof. First, by the product rule,

$$\mathbb{E}_{p(z)}\left(\psi_q(z) f(z) + f'(z)\right) = \int_z \frac{p(z)}{q(z)} \frac{\partial\left(q(z)f(z)\right)}{\partial z}\; dz$$

Now

$$\int \frac{q(z)}{q(z)} \frac{\partial(q(z)f(z))}{\partial z}\; dz = \int \frac{\partial}{\partial z}\left(q(z)f(z)\right)\; dz = q(\infty)f(\infty) - q(-\infty)f(-\infty) = 0$$

under some assumptions on q, f. □

Lemma 5.2 (Equivalence of score-matching and Stein discrepancy under specific $\mathcal{G}$)**.** *Let $\mathcal{G}$ be a class of functions of the form $g(z) = \psi_q(z)f(z) + f'(z)$, for f in some class of functions $\mathcal{F}$. Then*

$$\sup_{g\in\mathcal{G}} |\mathbb{E}_{p(z)} g(z)| = \sup_{f\in\mathcal{F}} |\mathbb{E}_{p(z)}\left(\psi_q(z)f(z) + f'(z)\right)| = \mathbb{E}_p\left(\psi_p - \psi_q\right)^2$$

Proof. By Lemma 5.1, $\mathbb{E}_{p(z)}\psi_p(z)f(z) + f'(z) = 0$. So

$$\begin{aligned}\sup_{g\in\mathcal{G}} |\mathbb{E}_{p(z)} g(z)| &= \sup_{f\in\mathcal{F}} |\mathbb{E}_{p(z)}\left(\psi_q(z)f(z) + f'(z) - \psi_p(z)f(z) - f'(z)\right)| \\ &= \sup_{f\in\mathcal{F}} |\mathbb{E}_p(\psi_q - \psi_p)f|\end{aligned}$$

And this is maximized when $f(z) = \psi_q(z) - \psi_p(z)$. □

Lemma 5.3. *Let $\mathcal{F}$ be a class of functions of the form (assuming the densities are positive everywhere):*

$$f(z) = \frac{p(z)}{q(z)} h(z) \tag{10}$$

for $h(z)$ in some class of functions $\mathcal{H}$. Then

$$\sup_{f\in\mathcal{F}} |\mathbb{E}_q O^p_{LS} f| = \sup_{h\in\mathcal{H}} |\mathbb{E}_p O^q_{LS} h|$$

Proof. This is clear if we use the representation in (8) and apply Lemma 5.2. That is,

$$\sup_{f\in\mathcal{F}}|\mathbb{E}_q O^p_{LS} f| = \sup_{f\in\mathcal{F}} \mathbb{E}_q(\psi_q - \psi_p)f = \sup_{h\in\mathcal{H}} \mathbb{E}_q \frac{p}{q}(\psi_q - \psi_p)h = \sup_{h\in\mathcal{H}} \mathbb{E}_p(\psi_q - \psi_p)h = \sup_{h\in\mathcal{H}}|\mathbb{E}_p O^q_{LS} h|$$

Alternatively (this is a sanity check):

$$\begin{aligned}
\sup_{f\in\mathcal{F}}|\mathbb{E}_q O^p_{LS} f| &= \sup_{f\in\mathcal{F}}|\mathbb{E}_q\left(\psi_p(z)f(z) + f'(z)\right)| \\
&= \sup_{h\in\mathcal{H}}\Big|\int q(z)\frac{p'(z)}{p(z)}\frac{p(z)}{q(z)}h(z) + \int q(z)\left(\frac{q(z)p'(z) - p(z)q'(z)}{q(z)^2}h(z) + \frac{p(z)}{q(z)}h'(z)\right)\Big| \\
&= \sup_{h\in\mathcal{H}}\Big|\int p'(z)h(z) + \int p'(z)h(z) - \int \frac{q'(z)}{q(z)}p(z)h(z) + \int p(z)h'(z)\Big| \\
&= \sup_{h\in\mathcal{H}}\Big|\int 2\left(p'(z)h(z) + p(z)h'(z)\right) - \int p(z)h'(z) - \int \frac{q'(z)}{q(z)}p(z)h(z)\Big| \\
&= \sup_{h\in\mathcal{H}}|\mathbb{E}_p\left(\psi_q(z)h(z) + h'(z)\right)|
\end{aligned}$$

since $p'(z)h(z) + p'(z)h(z) = \frac{\partial(p(z)h(z))}{\partial z}$, which has integral zero by Lemma 5.1. □

References

Hyvarinen, Aapo. 2005. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, **6**, 695–709.

Hyvarinen, Aapo. 2007. Some extensions of score matching. *Computational Statistics and Data Analysis*, 2499–2512.

Janofsky, Eric. 2015. *Exponential series approaches for nonparametric graphical models*. Ph.D. thesis, University of Chicago.

Lin, Lina, Drton, Mathias, & Shojaie, Ali. 2016. Estimation of high-dimensional graphical models using regularized score matching. *Electronic Journal of Statistics*, **10**(1), 806–854.

Ranganath, Rajesh, Tran, Dustin, Altosaar, Jaan, & Blei, David. 2016. Operator Variational Inference. *Pages 496–504 of:* Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., & Garnett, R. (eds), *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc.

Wang, Chong, & Blei, David M. 2013. Variational inference in nonconjugate models. *Journal of Machine Learning Research*, **14**, 1005–1031.