Stein's Unbiased Risk Estimator

Guy Lebanon

In this note we describe Stein's lemma and its use to derive Stein's unbiased risk estimator (SURE). We focus on exponential families as in [1] (though our exposition is informal and omits technical conditions).

Lemma 1. *Let $\theta, x \in \mathbb{R}^k$, and $S : \mathbb{R}^k \to \mathbb{R}$ satisfy*

$$\int |s(x)| \exp(\theta^\top x)\, dx < \infty, \qquad \int |s_i'(x)| \exp(\theta^\top x)\, dx < \infty \quad i = 1, \ldots, k. \tag{1}$$

Then

$$\theta_i \int_{\mathbb{R}^k} s(x) \exp(\theta^\top x)\, dx = -\int_{\mathbb{R}^k} s_i'(x) \exp(\theta^\top x)\, dx \qquad i = 1, \ldots, k. \tag{2}$$

Proof. Using integration by parts for the integral over x_1 we get

$$\begin{aligned} -\int_{\mathbb{R}^{k-1}} \theta_1 \int_{\mathbb{R}} s(x) \exp(\theta^\top x)\, dx_1 dx_2 \cdots dx_k &= \int_{\mathbb{R}^{k-1}} \int_{\mathbb{R}} s_i'(x) \exp(\theta^\top x)\, dx_1 \cdots dx_k \\ &\quad - \lim_{B \to \infty} \left[\int_{\mathbb{R}^{k-1}} s(x) \exp(\theta^\top x) dx_2 \cdots dx_k \right]_{x_1 = -B}^{x_1 = B} \\ &= \int_{\mathbb{R}^k} s_i'(x) \exp(\theta^\top x)\, dx_1 + 0 - 0 \end{aligned}$$

where the last terms on the right hand side above are zero due to (1). Repeating the argument above for $x_2, \ldots, x_k$ completes the proof. □

Corollary 1. *For the exponential family distribution $p_\theta(x) = h(x) \exp(\theta^\top x - \psi(x))$ and smooth and integrable functions $t : \mathbb{R}^k \to \mathbb{R}$, $r : \mathbb{R}^k \to \mathbb{R}^k$ we have the following vector equation and scalar equation*

$$\theta\ \mathsf{E}_{p_\theta}(t(X)) = -\mathsf{E}_{p_\theta}\left(\nabla t(X) + \frac{t(X) \nabla h(X)}{h(X)} \right) \tag{3}$$

$$\mathsf{E}_{p_\theta}(\theta^\top r(X)) = -\mathsf{E}_{p_\theta}\left(\nabla \cdot r(X) + \frac{r(X)^\top \nabla h(X)}{h(X)} \right). \tag{4}$$

where ∇f is the gradient $\nabla f = (f_1'(x), \ldots, f_k'(x))$ and $\nabla \cdot f$ is the divergence $\sum_i \partial f(x)/\partial x_i$.

Proof. For (3) we apply Lemma 1 with $s(x) = t(x)h(x)$ and note that $(th)_i' = (t_i' + h_i' t/h)h$. For (4) we apply (3) with $t = r_i$ for $i = 1, \ldots, k$ and sum. □

Corollary 2. *Using the notation of the previous corollary and assuming the intgral below converge absolutely*

$$\|\theta\|^2 = \mathsf{E}_{p_\theta} \frac{\nabla^2 h(X)}{h(X)} \tag{5}$$

where ∇^2 is the Laplacian $\nabla^2 f = \sum_i f_{ii}''$.

Proof. Use Lemma 1 twice (first time with $s = h$ and second time with $s = h'$) for $i = 1, \ldots, k$ and sum. □

Proposition 1. *Let $\delta : \mathbb{R}^k \to \mathbb{R}^k$ be an estimator of θ (based on a single sample $X \sim p_\theta$). Using the definitions above and assuming the derivatives and integrals below exist*

$$R(\theta, \delta) = \mathsf{E}_{p_\theta} \|\delta(X) - \theta\|^2 = \mathsf{E}_{p_\theta} \left(\|\delta(X)\|^2 - 2\nabla \cdot \delta(X) + 2\frac{\delta(X)^\top \nabla h(X)}{h(X)} + \frac{\nabla^2 h(X)}{h(X)} \right). \tag{6}$$

Proof. Note that

$$\mathsf{E}_{p_\theta} \|\delta(X) - \theta\|^2 = \mathsf{E}_{p_\theta} \|\delta(X)\|^2 + \|\theta\|^2 + 2\mathsf{E}_{p_\theta} \theta^\top \delta(X) \tag{7}$$

and use the previous two corollaries. □

Comments

1. Stein's unbiased risk estimator (SURE) is the integrand in the right hand side of (6). It does not depend on θ and so may be computed using the sample X and the exponential family structure. Since its expectation is the risk it is an unbiased estimator.

2. In some cases we are intrested in the risk difference between two estimators $R(\theta, \delta_1) - R(\theta, \delta_2)$. In this case SURE lacks the term $\|\theta\|^2$ that gets cancelled

$$R(\theta, \delta_1) - R(\theta, \delta_2) = \mathsf{E}_{p_\theta} \left(\|\delta_1(X)\|^2 - \|\delta_2(X)\|^2 + 2\nabla \cdot (\delta_1(X) - \delta_2(X)) + 2\frac{(\delta_1(X) - \delta_2(X))^\top \nabla h(X)}{h(X)} \right).$$

3. Note that we assumed above that X is continuous. For discrete exponential families there exists a SURE analog where the partial derivatives are replaced with the finite difference operator [1, Sec. 4.12].

4. The proposition above assumes that the estimator δ is based on a single sample $X \sim p_\theta$. But in most cases estimators are based on more than one observations i.e., $\delta_n(X^{(1)}, \ldots, X^{(n)})$, $X^{(1)}, \ldots, X^{(n)} \overset{\text{iid}}{\sim} p_\theta$. A simple way to extend SURE to this case is to consider $X^* = (X^{(1)}, \ldots, X^{(n)})$ sampled from

$$p_{\theta^*}(X^{(1)}, \ldots, X^{(n)}) = \left(\prod_i h(X^{(i)}) \right) \exp\left(\sum_{i=1}^n \theta^\top X^{(i)} - n\psi(\theta) \right), \quad \theta^* = (\theta, \ldots, \theta). \tag{8}$$

   Now apply SURE to $X^* \sim p_{\theta^*}$ and the estimator

$$\delta^* : \mathbb{R}^{kn} \to \mathbb{R}^{kn}, \quad \delta^*(X^{(1)}, \ldots, X^{(n)}) = (\delta_n(X^{(1)}, \ldots, X^{(n)}), \ldots, \delta_n(X^{(1)}, \ldots, X^{(n)}))$$

   and note that $R(\theta^*, \delta^*) = nR(\theta, \delta_n)$.

5. SURE has been used in many cases including as a substitute for cross validation in determining values of tuning parameters. On some occasions it performs better than cross validation and in other occasions worse. It has the distinct advantage of offering an analytic unbiased estimator (as opposed to an algorithmic procedure). Note however the requirement that δ has a closed form (rather than an algorithmic procedure) that is differentiable in X. Another popular use (and the context for its original appearance) is to prove the inadmissibility of the standard Gaussian means estimator and introduce the James-Stein estimator.

References

[1] L. D. Brown. *Fundamentals of Statistical Exponential Families*, volume 9 of *Lecture Notes-Monograph Series*. IMS Press, 1986.