Appendix for "Post-Inference Prior Swapping"

A. Details on the IS Example (Sec. 2.1)

Here we provide details on the IS example (for a normal π_f and Laplace π) given in Sec. 2.1.

We made the following statement: if $p_f(\theta|x^n) = \mathcal{N}(\theta|m, s^2)$, in order for $|\mu_h - \mathbb{E}_{p_f}[\hat{\mu}_h^{\text{IS}}]| < \delta$, we need

$$T \geq \exp\left\{\frac{1}{2s^2}(|\mu_h - m| - \delta)^2\right\}.$$

To show this, we first give an upper bound on the expected value of the maximum of T zero-mean s^2-variance Gaussian random variables. Let $\{\tilde{\theta}_t\}_{t=1}^T \sim g$, where $g(\theta) = \mathcal{N}(\theta|0, s^2)$, and let $Z = \max_t\{\tilde{\theta}_t\}_{t=1}^T$. Then, for some $b > 0$,

$$\exp\{b\mathbb{E}_g[Z]\} \leq \mathbb{E}_g[\exp\{bZ\}] = \mathbb{E}_g\left[\max_t\left\{\exp\{b\tilde{\theta}_t\}\right\}_{t=1}^T\right] \leq \sum_{t=1}^T \mathbb{E}_g\left[\exp\{b\tilde{\theta}_t\}\right] = T\exp\{b^2s^2/2\},$$

where the first inequality is due to Jensen's inequality, and the final equality is due to the definition of a Gaussian moment generating function. The above implies that

$$\mathbb{E}_g[Z] \leq \frac{\log T}{b} + \frac{bs^2}{2}.$$

Setting $b = \sqrt{\frac{2}{s^2}\log T}$, we have that

$$\mathbb{E}_g\left[\max_t\{\tilde{\theta}_t\}_{t=1}^T\right] = \mathbb{E}_g[Z] \leq s\sqrt{2\log T}.$$

However, note that for all $\{\tilde{\theta}_t\}_{t=1}^T$, and weights $\{w(\tilde{\theta}_t)\}_{t=1}^T$ (such that $\sum_{t=1}^T w(\tilde{\theta}_t) = 1$), the IS estimate $\hat{\mu}_h^{\text{IS}}$ for $h(\theta) = \theta$ must be less than or equal to $\max_t\{\tilde{\theta}_t\}_{t=1}^T$ (since the weighted average of $\{\tilde{\theta}_t\}_{t=1}^T$ cannot be larger than the maximum of this set). Therefore,

$$\mathbb{E}_g\left[\hat{\mu}_h^{\text{IS}}\right] \leq \mathbb{E}_g\left[\max_t\{\tilde{\theta}_t\}_{t=1}^T\right] \leq s\sqrt{2\log T},$$

and equivalently

$$T \geq \exp\left\{\frac{1}{2s^2}\mathbb{E}_g\left[\hat{\mu}_h^{\text{IS}}\right]^2\right\}.$$

In our example, we wanted the expected estimate to be within δ of μ_h, i.e. we wanted $|\mu_h - \mathbb{E}_g[\hat{\mu}_h^{\text{IS}}]| < \delta \iff \delta - \mu_h \leq \mathbb{E}_g[\hat{\mu}_h^{\text{IS}}] \leq \mu_h + \delta$, and therefore,

$$T \geq \exp\left\{\frac{1}{2s^2}\mathbb{E}_g\left[\hat{\mu}_h^{\text{IS}}\right]^2\right\} \geq \exp\left\{\frac{1}{2s^2}\left(\delta - \mu_h\right)^2\right\}.$$

Finally, notice that the original statement involved samples $\{\tilde{\theta}_t\}_{t=1}^T \sim p_f(\theta|x^n) = \mathcal{N}(m, s^2)$ (instead of from $g = \mathcal{N}(0, s^2)$). But this is equivalent to setting $p_f(\theta|x^n) = g(\theta)$, and shifting our goal so that we want $\delta - |\mu_h - m| \leq \mathbb{E}_{p_f}[\hat{\mu}_h^{\text{IS}}] \leq |\mu_h - m| + \delta$. This gives us the desired bound:

$$T \geq \exp\left\{\frac{1}{2s^2}\mathbb{E}_{p_f}\left[\hat{\mu}_h^{\text{IS}}\right]^2\right\} \geq \exp\left\{\frac{1}{2s^2}\left(\delta - |\mu_h - m|\right)^2\right\}.$$

B. Prior Swapping Pseudocode (for a false posterior PDF inference result $\tilde{p}_f(\theta)$)

Here we give pseudocode for the prior swapping procedure, given some false posterior PDF inference result $\tilde{p}_f(\theta)$, using the prior swap functions $p_s(\theta) \propto \frac{\tilde{p}_f(\theta)\pi(\theta)}{\pi_f(\theta)}$ and $\nabla_\theta \log p_s(\theta) \propto \nabla_\theta \log \tilde{p}_f(\theta) + \nabla_\theta \log \pi(\theta) - \nabla_\theta \log \pi_f(\theta)$, as described in Sec. 2.2.

In Alg. 2, we show prior swapping via the Metropolis-Hastings algorithm, which makes repeated use of $p_s(\theta)$. In Alg. 3 we show prior swapping via Hamiltonian Monte Carlo, which makes repeated use of $\nabla_\theta \log p_s(\theta)$. A special case of Alg. 3, which occurs when we set the number of simulation steps to $L = 1$ (in line 6), is prior swapping via Langevin dynamics.

Algorithm 2: Prior swapping via Metropolis-Hastings.

Input: Prior swap function $p_s(\theta)$, and proposal q.

Output: Samples $\{\theta_t\}_{t=1}^T \sim p_s(\theta)$ as $T \to \infty$.

1 Initialize θ_0. ▷ Initialize Markov chain.
2 **for** $t = 1, \ldots, T$ **do**
3 Draw $\theta_s \sim q(\theta_s \mid \theta_{t-1})$. ▷ Propose new sample.
4 Draw $u \sim \text{Unif}([0, 1])$.
5 **if** $u < \min\left\{1, \frac{p_s(\theta_s)q(\theta_t|\theta_s)}{p_s(\theta_t)q(\theta_s|\theta_t)}\right\}$ **then**
6 Set $\theta_t \leftarrow \theta_s$. ▷ Accept proposed sample.
7 **else**
8 Set $\theta_t \leftarrow \theta_{t-1}$. ▷ Reject proposed sample.

Algorithm 3: Prior swapping via Hamiltonian Monte Carlo.

Input: Prior swap function $p_s(\theta)$, its gradient-log $\nabla_\theta \log p_s(\theta)$, and step-size ϵ.

Output: Samples $\{\theta_t\}_{t=1}^T \sim p_s(\theta)$ as $T \to \infty$.

1 Initialize θ_0. ▷ Initialize Markov chain.
2 **for** $t = 1, \ldots, T$ **do**
3 Draw $r_t \sim \mathcal{N}(0, I)$.
4 Set $(\widetilde{\theta}_0, \widetilde{r}_0) \leftarrow (\theta_{t-1}, r_{t-1})$
5 Set $\widetilde{r}_0 \leftarrow \widetilde{r}_0 + \frac{\epsilon}{2}\nabla_\theta \log p_s(\widetilde{\theta}_0)$. ▷ Propose new sample (next 4 lines).
6 **for** $l = 1, \ldots, L$ **do**
7 Set $\widetilde{\theta}_l \leftarrow \widetilde{\theta}_{l-1} + \epsilon\widetilde{r}_{l-1}$.
8 Set $\widetilde{r}_l \leftarrow \widetilde{r}_{l-1} + \epsilon\nabla_\theta \log p_s(\widetilde{\theta}_l)$.
9 Set $\widetilde{r}_L \leftarrow \widetilde{r}_L + \frac{\epsilon}{2}\nabla_\theta \log p_s(\widetilde{\theta}_L)$.
10 Draw $u \sim \text{Unif}([0, 1])$.
11 **if** $u < \min\left\{1, \frac{p_s(\widetilde{\theta}_L)\widetilde{r}_L^\top \widetilde{r}_L}{p_s(\theta_{t-1})r_{t-1}^\top r_{t-1}}\right\}$ **then**
12 Set $\theta_t \leftarrow \widetilde{\theta}_L$. ▷ Accept proposed sample.
13 **else**
14 Set $\theta_t \leftarrow \theta_{t-1}$. ▷ Reject proposed sample.

C. Proofs of Theoretical Guarantees

Here, we prove the theorems stated in Sec. 2.3.

Throughout this analysis, we assume that we have T samples $\{\tilde{\theta}_t\}_{t=1}^{T_f} \subset \mathcal{X} \subset \mathbb{R}^d$ from the false-posterior $p_f(\theta|x^n)$, and that $b \in \mathbb{R}_+$ denotes the bandwidth of our semiparametric false-posterior density estimator $\tilde{p}_f^{sp}(\theta)$. Let Hölder class $\Sigma(2, L)$ on $\mathcal{X}$ be defined as the set of all $\ell = \lfloor 2 \rfloor$ times differentiable functions $f : \mathcal{X} \to \mathbb{R}$ whose derivative $f^{(l)}$ satisfies

$$|f^{(\ell)}(\theta) - f^{(\ell)}(\theta')| \leq L\,|\theta - \theta'|^{2-\ell} \quad \text{for all } \theta, \theta' \in \mathcal{X}.$$

Let the class of densities $\mathcal{P}(2, L)$ be

$$\mathcal{P}(2, L) = \left\{ f \in \Sigma(2, L) \;\middle|\; f \geq 0, \int f(\theta) d\theta = 1 \right\}.$$

Let data $x^n = \{x_1, \ldots, x_n\} \subset \mathcal{Y} \subset \mathbb{R}^p$, let $\mathcal{Z} \subset \mathcal{Y}$ be any set such that $x^n \subset \mathcal{Z}$, and let $\mathcal{F}_{\mathcal{Z}}(L)$ denote the set of densities $p : \mathcal{Y} \to \mathbb{R}$ that satisfy

$$|\log p(x) - \log p(x')| \leq L|x - x'|, \quad \text{for all } x, x' \in \mathcal{Z}.$$

In the following theorems, we assume that the false-posterior density $p_f(\theta|x^n)$ is bounded, i.e. that there exists some $B > 0$ such that $p_f(\theta|x^n) \leq B$ for all $\theta \in \mathbb{R}^d$; that the prior swap density $p_s(\theta) \in \mathcal{P}(2, L)$; and that the model family $p(x^n|\theta) \in \mathcal{F}_{\mathcal{Z}}(L)$ for some $\mathcal{Z}$.

Theorem 2.1. *For any $\alpha = (\alpha_1, \ldots, \alpha_k) \subset \mathbb{R}^p$ and $k > 0$ let $\tilde{p}_f^{\alpha}(\theta)$ be defined as in Eq. (8). Then, there exists $M > 0$ such that $\frac{p_f(\theta|x^n)}{\tilde{p}_f^{\alpha}(\theta)} < M$, for all $\theta \in \mathbb{R}^d$.*

Proof. To prove that there exists $M > 0$ such that $\frac{p_f(\theta|x^n)}{\tilde{p}_f^{\alpha}(\theta)} < M$, note that the false posterior can be written

$$p_f(\theta|x^n) = \frac{1}{Z_1}\pi_f(\theta)\prod_{i=1}^{n} L(\theta|x_i) = \frac{1}{Z_1}\pi_f(\theta)\prod_{i=1}^{n} p(x_i|\theta),$$

and the parametric estimate $\tilde{p}_f^{\alpha}(\theta)$ is defined to be

$$\tilde{p}_f^{\alpha}(\theta) = \frac{1}{Z_2}\pi_f(\theta)\prod_{j=1}^{k} p(\alpha_j|\theta)^{n/k}.$$

Let $d = \max_{i,j} |x_i - \alpha_j|$. For any $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, k\}$,

$$|\log p(x_i|\theta) - \log p(\alpha_j|\theta)| \leq Ld \implies \left|\log \frac{p(x_i|\theta)}{p(\alpha_j|\theta)}\right| \leq Ld,$$

and

$$\exp\left\{\log \frac{p(x_i|\theta)}{p(\alpha_j|\theta)}\right\} \leq \exp\left\{\left|\log \frac{p(x_i|\theta)}{p(\alpha_j|\theta)}\right|\right\} \leq \exp\{Ld\} \implies \frac{p(x_i|\theta)}{p(\alpha_j|\theta)} \leq \exp\{Ld\}.$$

Therefore

$$\frac{p_f(\theta|x^n)}{\tilde{p}_f^{\alpha}(\theta)} \leq \frac{Z_2}{Z_1}\frac{\prod_{i=1}^{n} p(x_i|\theta)}{\prod_{j=1}^{k} p(\alpha_j|\theta)^{n/k}} \leq \frac{Z_2}{Z_1}\exp\{nLd\} = M.$$

□

Corollary 2.1.1. *For* $\{\theta_t\}_{t=1}^T \sim p_s^\alpha(\theta) \propto \frac{\tilde{p}_f^\alpha(\theta)\pi(\theta)}{\pi_f(\theta)}$, $w(\theta_t) = \frac{p_f(\theta_t|x^n)}{\tilde{p}_f^\alpha(\theta_t)} \left(\sum_{r=1}^T \frac{p_f(\theta_r|x^n)}{\tilde{p}_f^\alpha(\theta_r)}\right)^{-1}$, *and test function that satisfies* $Var_p[h(\theta)] < \infty$, *the variance of IS estimate* $\hat{\mu}_h^{PSis} = \sum_{t=1}^T h(\theta_t)w(\theta_t)$ *is finite.*

Proof. This follows directly from the sufficient conditions for finite variance IS estimates given by (Geweke, 1989), which we have proved are satisfied for $\hat{\mu}_h^{\text{PSis}}$ in Theorem 2.1. □

Theorem 2.2. *Given false posterior samples* $\{\tilde{\theta}_t\}_{t=1}^{T_f} \sim p_f(\theta|x^n)$ *and* $b \asymp T_f^{-1/(4+d)}$, *the estimator* p_s^{sp} *is consistent for* $p(\theta|x^n)$, *i.e. its mean-squared error satisfies*

$$\sup_{p(\theta|x^n)\in\mathcal{P}(2,L)} \mathbb{E}\left[\int \left(p_s^{sp}(\theta) - p(\theta|x^n)\right)^2 d\theta\right] < \frac{c}{T_f^{4/(4+d)}}$$

for some $c > 0$ *and* $0 < b \leq 1$.

Proof. To prove mean-square consistency of our semiparametric prior swap density estimator p_s^{sp}, we give a bound on the mean-squared error (MSE), and show that it tends to zero as we increase the number of samples T_f drawn from the false-posterior. To prove this, we bound the bias and variance of the estimator, and use this to bound the MSE. In the following, to avoid cluttering notation, we will drop the subscript p_f in $\mathbb{E}_{p_f}[\cdot]$.

We first bound the bias of our semiparametric prior swap estimator. For any $p(\theta|x^n) \in \mathcal{P}(2, L)$, we can write the bias as

$$\begin{aligned}
\left|\mathbb{E}\left[p_s^{sp}(\theta)\right] - p(\theta|x^n)\right| &= c_1 \left|\mathbb{E}\left[\tilde{p}_f^{sp}(\theta)\frac{\pi(\theta)}{\pi_f(\theta)}\right] - p_f(\theta|x^n)\frac{\pi(\theta)}{\pi_f(\theta)}\right| \\
&= c_2 \left|\frac{\pi(\theta)}{\pi_f(\theta)}\mathbb{E}\left[\tilde{p}_f^{sp}(\theta)\right] - p_f(\theta|x^n)\right| \\
&= c_3 \left|\mathbb{E}\left[\tilde{p}_f^{sp}(\theta)\right] - p_f(\theta|x^n)\right| \\
&\leq ch^2
\end{aligned}$$

for some $c > 0$, where we have used the fact that $\left|\mathbb{E}\left[\tilde{p}_f^{sp}(\theta)\right] - p_f(\theta|x^n)\right| \leq \tilde{c}h^2$ for some $\tilde{c} > 0$ (given in (Hjort & Glad, 1995; Wasserman, 2006)).

We next bound the variance of our semiparametric prior swap estimator. For any $p(\theta|x^n) \in \mathcal{P}(2, L)$, we can write the variance of our estimator as

$$\begin{aligned}
\text{Var}\left[p_s^{sp}(\theta)\right] &= c_1 \text{Var}\left[\tilde{p}_f^{sp}(\theta)\frac{\pi(\theta)}{\pi_f(\theta)}\right] \\
&= \frac{\pi(\theta)^2}{\pi_f(\theta)^2}\text{Var}\left[\tilde{p}_f^{sp}(\theta)\right] \\
&\leq \frac{c}{T_f h^d}
\end{aligned}$$

for some $c > 0$, where we have used the facts that $\text{Var}\left[\tilde{p}_f^{sp}(\theta)\right] \leq \frac{c}{Th^d}$ for some $c > 0$ and $\mathbb{E}\left[\tilde{p}_f^{sp}(\theta)\right]^2 \leq \tilde{c}$ for some $\tilde{c} > 0$ (given in (Hjort & Glad, 1995; Wasserman, 2006)). Next, we will use these two results to bound the mean-squared error of our semiparametric prior swap estimator, which shows that it is mean-square consistent.

We can write the mean-squared error as the sum of the variance and the bias-squared, and therefore,

$$\begin{aligned}
\mathbb{E}\left[\int \left(p_s^{sp}(\theta) - p(\theta|x^n)\right)^2 d\theta\right] &\leq c_1 h^2 + \frac{c_2}{Th^d} \\
&= \frac{c}{T_f^{4/(4+d)}}
\end{aligned}$$

for some $c > 0$, using the fact that $h \asymp T_f^{-1/(4+d)}$. □

D. Further Empirical Results

Here we show further empirical results on a logistic regression model with hierarchical target prior given by $\pi = \mathcal{N}(0, \alpha^{-1}I)$, $\alpha \sim$ Gamma$(\gamma, 1)$. We use synthetic data so that we are able to compare the timing and posterior error of different methods as we tune n and d.

In this experiment, we assume that we are given samples from a false posterior $p_f(\theta|x^n)$, and we want to most-efficiently compute the target posterior under prior $\pi(\theta)$. In addition to the prior swapping methods, we can run standard iterative inference algorithms, such as MCMC or variational inference (VI), on the target posterior (initializing them, for example, at the false posterior mode) as comparisons. The following experiments aim to show that, once the data size n grows large enough, prior swapping methods become more efficient than standard inference algorithms. They also aim to show that the held-out test error of prior swapping matches that of these standard inference algorithms. In these experiments, we also add a prior swap method called *prior swapping VI*; this method involves making a VI approximation to $p_f(\theta|x^n)$, and using it for $\tilde{p}_f(\theta)$. Prior swapping VI allows us to see whether the test error is similar to standard VI inference algorithms, which compute some approximation to the posterior. Finally, we show results over a range of target prior hyperparameter values γ to show that prior swapping maintains accuracy (i.e. has a similar error as standard inference algorithms) over the full range.

We show results in Fig. 6. In (a) and (b) we vary the number of observations (n=10-120,000) and see that prior swapping has a constant wall time while the wall times of both MCMC and VI increase with n. In (b) we see that the prior swapping methods achieve the same test error as the standard inference methods. In (c) and (d) we vary the number of dimensions (d=1-40). In this case, all methods have increasing wall time, and again the test errors match. In (e), (f), and (g), we vary the prior hyperparameter (γ=1-1.05). For prior swapping, we infer a single $\tilde{p}_f(\theta)$ (using $\gamma = 1.025$) with both MCMC and VI applied to $p_f(\theta|x^n)$, and compute *all other* hyperparameter results using this $\tilde{p}_f(\theta)$. This demonstrates that prior swapping can quickly infer correct results over a range of hyperparameters. Here, the prior swapping semiparametric method matches the test error of MCMC slightly better than the parametric method.

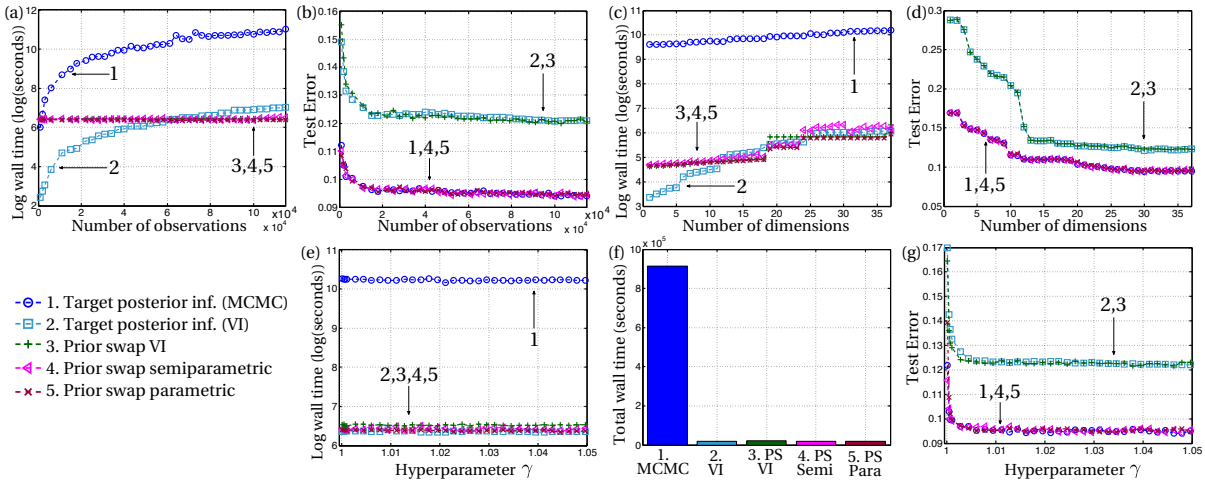

Figure 6. Bayesian hierarchical logistic regression: (a-b) Wall time and test error comparisons for varying data size n. As n is increased, wall time remains constant for prior swapping but grows for standard inference methods. (c-d) Wall time and test error comparisons for varying model dimensionality d. (e-g) Wall time and test error comparisons for inferences on a set of prior hyperparameters $\gamma \in [1, 1.05]$. Here, a single false posterior $\tilde{p}_f(\theta)$ (computed at $\gamma = 1.025$) is used for prior swapping on all other hyperparameters.