A Supplementary Material - Proofs

A.1 Proof of Theorem 3

Proof of Theorem 3. From Theorem 1 we immediately get the inequalities

$$-g(x) \leq -g^{\lambda}(x) \tag{15}$$

$$-g^{\lambda}(x) \leq -g(x) + \frac{L^2\lambda}{2}, \tag{16}$$

and thus also

$$\int \exp(-g^{\lambda}(z))dz \geq \int \exp(-g(z))dz = 1 \tag{17}$$

and

$$\int \exp(-g^{\lambda}(z))dz \leq \int \exp(-g(z))\exp(L^2\lambda/2)dz = \exp(L^2\lambda/2). \tag{18}$$

Let $f \geq 0$, then

$$\begin{aligned}
\mathbb{E}_{\pi^{\lambda}}(f) &= \int f(x)\frac{\exp(-g^{\lambda}(x))}{\int \exp(-g^{\lambda}(z))dz}dx \\
&\overset{(17)}{\leq} \int f(x)\exp(-g^{\lambda}(x))dx \\
&\overset{(16)}{\leq} \exp(L^2\lambda/2)\int f(x)\exp(-g(x))dx = \exp(L^2\lambda/2)\mathbb{E}_{\pi}(f).
\end{aligned}$$

Similarly, again for $f \geq 0$,

$$\begin{aligned}
\mathbb{E}_{\pi^{\lambda}}(f) &= \int f(x)\frac{\exp(-g^{\lambda}(x))}{\int \exp(-g^{\lambda}(z))dz}dx \\
&\overset{(18)}{\geq} \exp(-L^2\lambda/2)\int f(x)\exp(-g^{\lambda}(x))dx \\
&\overset{(15)}{\geq} \exp(-L^2\lambda/2)\int f(x)\exp(-g(x))dx \\
&= \exp(-L^2\lambda/2)\mathbb{E}_{\pi}(f).
\end{aligned}$$

In summary, for any non-negative f, we have

$$\exp(-L^2\lambda/2)\mathbb{E}_\pi(f) \le \mathbb{E}_{\pi^\lambda}(f) \le \exp(L^2\lambda/2)\mathbb{E}_\pi(f). \tag{19}$$

Subtracting $\mathbb{E}_\pi(f) \ge 0$ from these inequalities lets us derive

$$\begin{aligned} -(\exp(L^2\lambda/2)-1)\mathbb{E}_\pi(f) &= -\max\{\exp(L^2\lambda/2)-1, 1-\exp(-L^2\lambda/2)\}\mathbb{E}_\pi(f) \\ &= \min\{1-\exp(L^2\lambda/2), \exp(-L^2\lambda/2)-1\}\mathbb{E}_\pi(f) \\ &\le (\exp(-L^2\lambda/2)-1)\mathbb{E}_\pi(f) \\ &\overset{(19)}{\le} \mathbb{E}_{\pi^\lambda}(f) - \mathbb{E}_\pi(f) \\ &\overset{(19)}{\le} (\exp(L^2\lambda/2)-1)\mathbb{E}_\pi(f) \\ &\le \max\{\exp(L^2\lambda/2)-1, 1-\exp(-L^2\lambda/2)\}\mathbb{E}_\pi(f) \\ &= (\exp(L^2\lambda/2)-1)\mathbb{E}_\pi(f), \end{aligned} \tag{20}$$

and therefore

$$|\mathbb{E}_{\pi^\lambda}(f) - \mathbb{E}_\pi(f)| \le (\exp(L^2\lambda)-1)\mathbb{E}_\pi(f) \tag{21}$$

holds for any non-negative f.

For general f, we consider the standard decomposition $f = f^+ - f^-$ with $f^+ \ge 0$ and $f^- \ge 0$. Then $|f| = f^+ + f^-$, and as

$$\begin{aligned} |\mathbb{E}_{\pi^\lambda}(f) - \mathbb{E}_\pi(f)| &= |\mathbb{E}_{\pi^\lambda}(f^+) - \mathbb{E}_\pi(f^+) - [\mathbb{E}_{\pi^\lambda}(f^-) - \mathbb{E}_\pi(f^-)]| \\ &\le |\mathbb{E}_{\pi^\lambda}(f^+) - \mathbb{E}_\pi(f^+)| + |\mathbb{E}_{\pi^\lambda}(f^-) - \mathbb{E}_\pi(f^-)| \\ &\overset{(21)}{\le} (\exp(L^2\lambda)-1)\mathbb{E}_\pi(f^+) + (\exp(L^2\lambda)-1)\mathbb{E}_\pi(f^-) \\ &= (\exp(L^2\lambda)-1)\mathbb{E}_\pi(|f|), \end{aligned}$$

we have proved the first bound in the Theorem.

Since we can exchange the roles of π and π_λ in (19), we can follow the same chain of arguments to also get

$$|\mathbb{E}_{\pi^\lambda}(f) - \mathbb{E}_\pi(f)| \leq (\exp(L^2\lambda/2) - 1)\mathbb{E}_{\pi^\lambda}(|f|).$$

If $g = g_1 + g_2$ with Lipschitz-continuous g_1 and differentiable, but not necessarily Lipschitz-continuous, g_2, one takes the MYE of g_1 and notes that 15 and 16 hold for g_1. Adding g_2 on both sides of the inequality shows that these inequalities remain true for g such that the proof still holds. □

A.2 Proof of Lemma 1

Proof. The case $\lambda_1 = \lambda_2$ is trivial so assume $\lambda_1 < \lambda_2$.

Firstly recall that for convex g any MYE is also convex. Further note that g^{λ_2} is a Moreau-Yosida envelope for g^{λ_1}, with $g^{\lambda_2} = (g^{\lambda_1})^{\lambda_2 - \lambda_1}$ [4, Proposition 12.22 (ii)].

We may thus define $h = g^{\lambda_1}$, $\lambda = \lambda_2 - \lambda_1$, such that the statement of the lemma is equivalent to

Lemma (Equivalent Formulation of Lemma 1)**.** *For any convex and differentiable function $h : \mathcal{X} \to]-\infty, \infty]$, and for any $\lambda > 0$, the Moreau-Yosida envelope h^λ satisfies*

$$\|\nabla h(x)\| \geq \|\nabla h^\lambda(x)\| \quad \forall x \in \mathcal{X}.$$

We define $p = \text{prox}_h^\lambda(x)$. By theorem 2, $\nabla h^\lambda(x) = (x - p)/\lambda$; and by convexity (and differentiability) of h, we have for any $x \in \mathcal{X}$:

$$\begin{aligned}
0 &\leq \langle \nabla h(p) - \nabla h(x), p - x \rangle \\
&= \langle \nabla h(p) - \nabla h(x), -\lambda \nabla h(p) \rangle \\
&= -\lambda \|\nabla h(p)\|^2 + \langle \nabla h(x), \nabla h(p) \rangle \\
&\leq -\lambda \|\nabla h(p)\|^2 + \frac{\lambda}{2}\|\nabla h(x)\|^2 + \frac{\lambda}{2}\|\nabla h(p)\|^2 \\
&= \frac{\lambda}{2}\|\nabla h(x)\|^2 - \frac{\lambda}{2}\|\nabla h(p)\|^2,
\end{aligned}$$

where the first inequality is a necessary and sufficient condition for convexity of a differentiable function, and the last inequality follows from Young's inequality as $\langle x, y \rangle \leq \|x\|^2/2 + \|y\|^2/2$.

$$\|\nabla h(x)\|^2 \geq \|\nabla h(p)\|^2 \overset{(2)}{=} \|\frac{1}{\lambda}(x - p)\| = \|\nabla h^\lambda(x)\|$$

as required. The last equality is given by Theorem 2. □

A.3 Proof of Lemma 2

Proof. Invariance follows if

$$\int \mathcal{L}_{ZZ} f(x, v) \pi(dx) p(dv) = \int_{A_0} \mathcal{L}_{ZZ} f(x, v) \pi(dx) p(dv) + \int_{A_0^c} \mathcal{L}_{ZZ} f(x, v) \pi(dx) p(dv) = 0$$

for any $f \in D(\mathcal{L}_{ZZ})$, the domain of $\mathcal{L}_{ZZ}$ [28]. If the prior is differentiable such that π has a differentiable density, A_0 is empty and the proof is directly as in [7, Theorem 2.2]. If the prior is non-differentiable, A_0 is non-empty but a null-set under n-dimensional Lebesgue measure. Since π and $p(dv)$ are absolutely continuous with respect to Lebesgue measure, it follows that the first integral is zero. Invariance then again follows directly from [7, Theorem 2.2]. □

A.4 Proof of Lemma 3

Proof. As in Lemma 2, invariance with respect to the joint distribution of (x, v) follows if

$$\int \mathcal{L}_{BPS} f(x,v)\pi(dx)p(dv) = \int_{A_0} \mathcal{L}_{BPS} f(x,v)\pi(dx)p(dv) + \int_{A_0^c} \mathcal{L}_{BPS} f(x,v)\pi(dx)p(dv) = 0$$

for any $f \in D(\mathcal{L}_{BPS})$. Similarly to the proof of Lemma 2, the proof of [9, Proposition 1] applies directly under absolute continuity of π and $p(v)$ with respect to Lebesgue measure. □

A.5 Discretizing the Underdamped Langevin Dynamics

We implement the discretization used in [40]. If the current position and velocity are (x_t, v_t), the next iteration is given by

$$\begin{cases} x_{t+1} = x_t + \frac{1-\beta}{\gamma} v_t - \frac{1}{\gamma}(\nu - \frac{1-\beta}{\gamma\xi})\nabla U^\lambda(x_t) + W_x \\ v_{t+1} = \beta v_t - \frac{1-\beta}{\gamma\xi}\nabla U^\lambda(x_t) + W_v, \end{cases}$$

where $\nu = t_{n+1} - t_n$ is the step size, $\beta = \exp(-\gamma\xi\nu)$, and $(W_x, W_v) \sim \mathcal{N}(0, \Sigma)$ is Gaussian noise with covariance

$$\Sigma = \begin{pmatrix} \frac{1}{\gamma}\left(2\nu - \frac{3}{\gamma\xi} + \frac{4\beta}{\gamma\xi} - \frac{\beta^2}{\gamma\xi}\right) I_{d\times d} & \frac{1+\beta^2-2\beta}{\gamma\xi} I_{d\times d} \\ \frac{1+\beta^2-2\beta}{\gamma\xi} I_{d\times d} & \frac{1-\beta^2}{\xi} I_{d\times d} \end{pmatrix}.$$

All the experiments in this paper were run with $\gamma = 2$, $L = 1\lambda$, and $\nu = 2\lambda$, where λ is the tightness parameter of the respective MYE.

A.6 Hamiltonian Bouncy Particle Sampler

An alternative specification for the dynamics of the BPS was introduced in [56], which we will now detail. Consider the Hamiltonian of both the target variable and the velocity

$H(x,v)$

$$H(x,v) = U(x) + \log p(v) = -\ell(x) - \log \pi_0(x) - \frac{1}{2}v^t v + c,$$

where c is some constant we will suppress from now on. For some spherical potential $V(x) = \frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)$, consider now the augmented Hamiltonian

$$H(x,v) = \underbrace{-\ell(x) - \log \pi_0 - V(x)}_{\hat{U}(x)} + \underbrace{V(x) - \frac{1}{2}v^t}_{\hat{H}(x,v)},$$

which naturally can be broken into two parts. For the new system consisting of the latter two terms, the dynamics of the Hamiltonian $\hat{H}$ are available in closed form since the system of ODEs

$$\frac{\partial v_t}{\partial t} = -\nabla_{x_t}\hat{H}(x_t, v_t) = -\Sigma^{-1}(x_t - \mu)$$
$$\frac{\partial x_t}{\partial t} = \nabla_{v_t}\hat{H}(x_t, v_t) = v_t$$

can be solved explicitly for any μ and Σ. For the first three terms, we note that if the model under consideration has a Gaussian component in x then the spherical potential can be chosen to equal this energy function. For example, if π_0 is Gaussian, setting $V(x) = -\log \pi_0(x)$ reduces the Hamiltonian to only depend on the likelihood. As shown in [56], the resulting Hamiltonian BPS with rates and reflection operator given by

$$\hat{\rho}(t) = \max\{0, \langle v_t, \nabla \hat{U}(x_t)\rangle\}$$
$$\hat{\mathfrak{R}}_x v = v - 2\frac{\langle v, \nabla \hat{U}(x)\rangle}{\|\nabla \hat{U}(x)\|^2}\nabla \hat{U}(x),$$

and flow determined by $\hat{H}(x,v)$, has $\pi(x)v(x)$ as invariant distribution. It is clear that this is valid under any choice of μ and Σ, in particular, the Hamiltonian BPS can be localized if a factor decomposition is explicitly available.

B Supplementary Material - Further Experiments

B.1 Anisotropic Gaussian

To assess how the different algorithms compare on a *strongly* log-concave example, we repeated Example 4.1 with a centered Gaussian distribution, as in [9, Example 4.4]. The 100-dimensional distribution has a diagonal covariance matrix with $\Sigma_{i,i} = 1/i^2$. Following the guidance in [27], we picked $\lambda = 1/10^4$, as this is the Lipschitz constant of the log-gradient, and chose a step size $\delta/2 = \lambda$ for MY-ULA, and $\delta = 0.005$ for SK-ROCK. For pMALA, we set $\delta = 2\lambda$, and then chose $\lambda = 3 \times 10^{-5}$, giving an acceptance probability of around 60%. The results are summarised in Figure 5. The BPS is again run in its global form, a localized version thereof would improve performance. Estimates of the effective sample size per second are summarised in Table 3.

Algorithm	MY-ULA	MY-UULA	SK-ROCK	pMALA	BPS	ZZS
$\beta = 1$	2.00	1.39	2.28	1.27	1.17	2.48
$\beta = 100$	1774.74	1257.58	195.86	551.61	809.93	312.29

Table 3: Effective sample size per second for the different algorithm when targeting an anisotropic Gaussian distribution. Recall that the first three algorithms are asymptotically biased, while the last three are asymptotically exact.

B.2 Nuclear-norm models for low-rank matrix estimation

As a final illustration of our methods performance in exact sampling, we consider a nuclear-norm model example taken from [45]. Let $x \in \mathbb{R}^{n\times n}$ be an unknown low-rank matrix, and

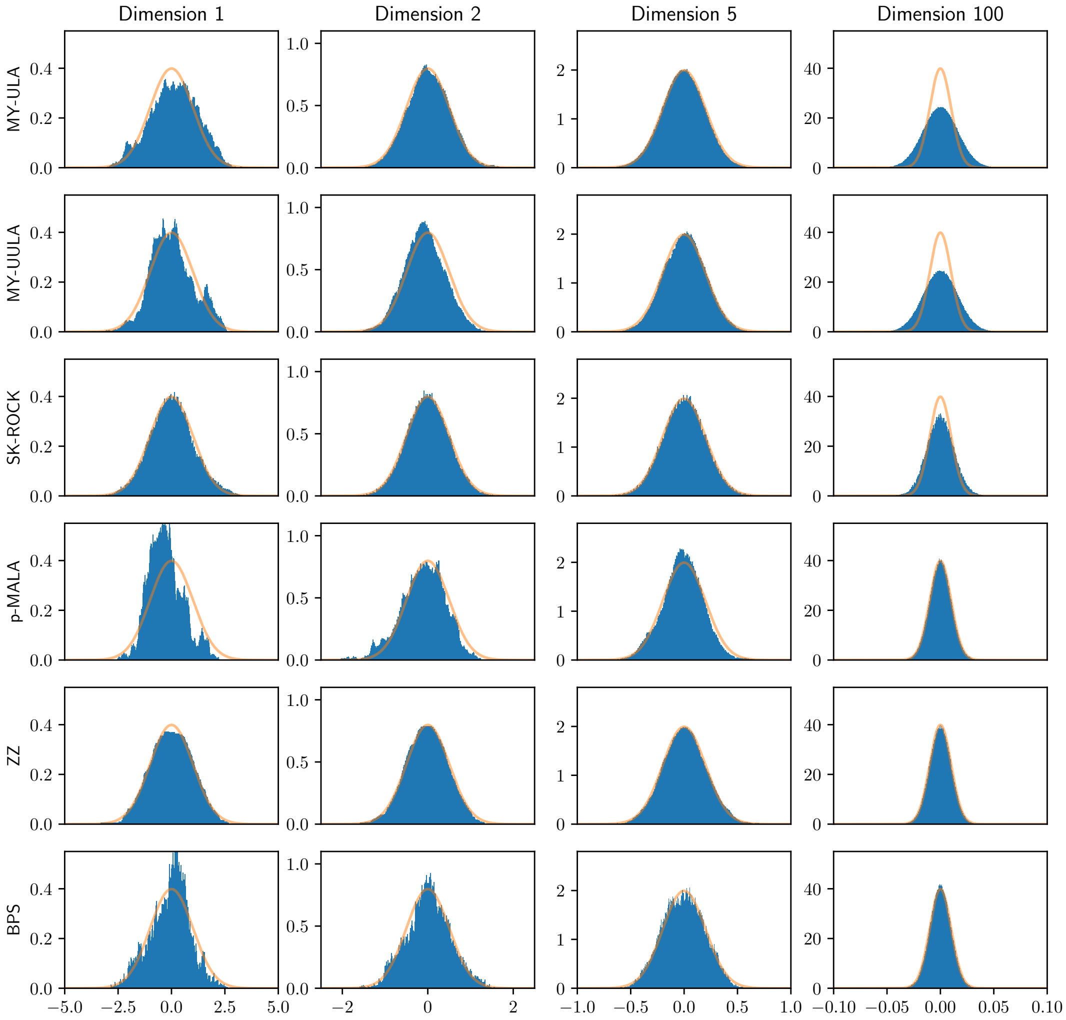

Figure 5: All algorithms are targeting a 100-dimensional anisotropic Gaussian distribution. The first three rows correspond to the approximate algorithms, and none of them manage to fully capture the narrowest component. The ZZS perfectly captures the last component, and shows good results in the first component. The BPS (in its global form) mixes slowly in the first component, but well in the last. All algorithms were given the same computational budget for a fair comparison.

let observations be noisy measurements thereof: $y = x + \xi$, where the entries of ξ are i.i.d. $N(0, \sigma^2)$. We assume that x is a low-rank matrix, and our aim is to sample from the posterior distribution of x given by

$$\pi(x) \propto \exp\left(-\frac{1}{2\sigma^2}\|x - y\|_F - \alpha\|x\|_*\right), \tag{22}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\|\cdot\|_*$ denotes the nuclear norm which favors low-rank matrices and penalizes high-rank ones. Conveniently, the proximal operator of the nuclear norm is available in closed form: Let $x = Q\Sigma V^T$ be the singular value decomposition of x, with $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$. Then the proximal operator is given by

$$\operatorname{prox}^{\lambda}_{\alpha\|\cdot\|_*}(x) = Q\operatorname{diag}\left(\operatorname{sgn}(\sigma_1)\max(|\sigma_1| - \alpha\lambda, 0), \ldots, \operatorname{sgn}(\sigma_n)\max(|\sigma_n| - \alpha\lambda, 0)\right)V^T,$$

i.e., one applies the soft thresholding operator to the singular values of x. We can thus efficiently compute the gradient to use in the Langevin-based samplers,

$$\nabla U^{\lambda}(x) = \frac{1}{\sigma^2}(x - y) + \frac{1}{\lambda}\Big(x - \operatorname{prox}^{\lambda}_{\alpha\|\cdot\|_*}(x)\Big). \tag{23}$$

We generated y by adding Gaussian noise to a matrix x^{true} with entries $x^{\text{true}}_{i,j} \in \{0, 0.7, 1\}$. The matrix x^{true} is visually a checkerboard with white, grey, or black checks.

We set $\lambda = \sigma^2$. The step size for MY-ULA is set to $\delta = 2\lambda$. A particular issue for the BPS in this model is the lack of factor decomposition due both to non-linearity of the nuclear norm and the proximal operator, which prevents us from using a localized, and therefore faster, version of the BPS. In an attempt to mitigate the resulting debilitated dynamics, we note that the likelihood in this case is equivalent to a isotropic Gaussian distribution in x as well. Defining an auxiliary potential by $V(x|y) = \|x - y\|^2/2$, we propose to generate dynamics according to the Hamiltonian flow (see A.6) corresponding to $(\dot{x}, \dot{v}) = (v_t, -(x_t - y)/\sigma^2)$, which has the explicit solution

$$\begin{pmatrix} x_t \\ v_t \end{pmatrix} = \begin{pmatrix} v_0 \sin\left(\frac{t}{\sigma}\right)\sigma + (x_0 - y)\cos\left(\frac{t}{\sigma}\right) + y \\ -(x_0 - y)\sin\left(\frac{t}{\sigma}\right) + v_0 \cos\left(\frac{t}{\sigma}\right) \end{pmatrix}.$$

By this choice of V it follows that the gradient employed in the rate and reflection operator subsequently is

$$\nabla \hat{U}^{\lambda}(x) = \frac{1}{\lambda}\Big(x - \operatorname{prox}^{\lambda}_{\alpha\|\cdot\|_*}(x)\Big). \tag{24}$$

Figure 6 shows the mean squared error between the posterior mean estimate of the respective algorithms, as calculated every second, and the 'true' posterior mean, as estimated by a very long run using an asymptotically unbiased algorithm. All algorithms are started at the same point, not too far away from the region of high probability. One can see that while MY-ULA quickly gives good estimates, the second-order scheme MY-UULA quickly yields better estimates. Interestingly, SK-ROCK performs worse here. The BPS does not yield any useful estimates in reasonable time, but after a while the HBPS gives the second best results. For completeness, we note that the Zig-Zag Sampler is not able to computationally compete with any of the other methods, as a single reflection requires the evaluation of the full gradient, which is prohibitively expensive. We also estimated the slowest and fastest mixing components of the checkerboard by estimating the sample covariance matrix during a long run of an exact sampler, and taking the first and last eigenvector thereof as the direction where the chain mixes slowest, and fastest, respectively. The autocorrelation plots for these components are shown in the second and third panel of Figure 6.

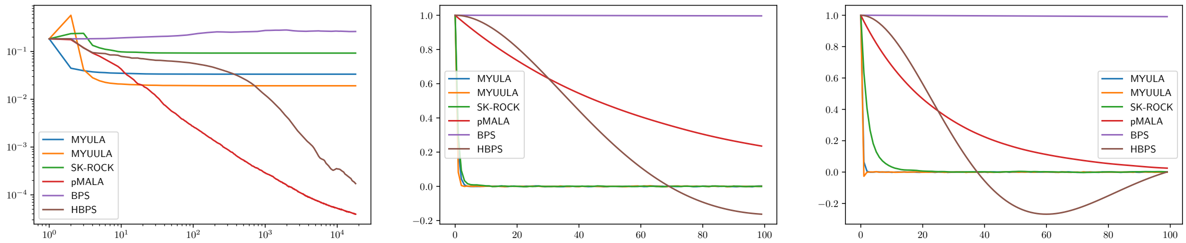

Figure 6: Results from the nuclear norm example. Left: MSE over time, for the different algorithms, run for half an hour each, on a log-log-scale. Middle: Autocorrelation for the slowest component, sample number adjusted for a fair comparison. Right: Autocorrelation for the fastest component, sample number adjusted for a fair comparison.

B.3 Image Deblurring

Uncertainty quantification in images is generally a challenging computational problem, with samples from the posterior used to estimate credible intervals or provide model comparisons. We focus on a purely illustrative example involving the total variation prior similar to [27, Example 4.1.2]. Let $x \in \mathbb{R}^{n_1 \times n_2}$ be an image which we observe through $y = Hx + \xi$, where H is a blurring operator that blurs a pixel $x_{i,j}$ uniformly with its closest neighbours (5×5 patch), and $\xi \sim N(0, \sigma^2 I_{n_1 \times n_2})$. The log-prior is proportional to $-TV(x) = -\alpha \|\nabla_D x\|_1$, where ∇_D is the two-dimensional discrete gradient operator as defined in [16], and α is a fixed parameter. The application of the TV prior is common in a wide array of imaging applications, as it emphasizes smooth surfaces bounded by distinct edges. As the authors of [27] we chose the 256×256 "boat" test image, and set $\alpha = 0.03$, $\sigma = 0.47$. The posterior is given by

$$\pi(x) \propto \exp\left(-\frac{1}{2\sigma^2}\|Hx - y\|_2^2 - \alpha TV(x)\right). \tag{25}$$

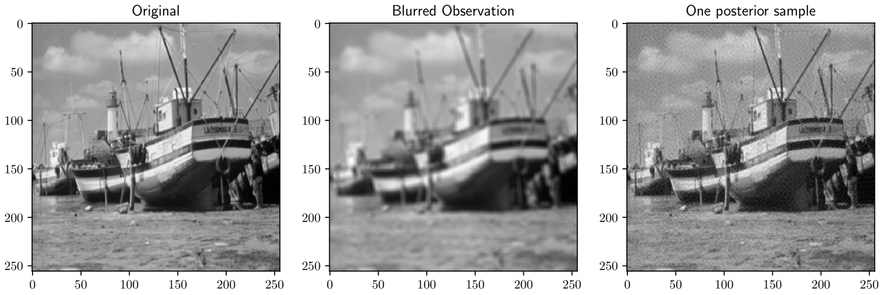

Figure 7: Left: The original 256 × 256 image. Center: The image after the application of the uniform blur operator. Right: A representative sample from the posterior distribution given in equation (25), obtained using the LBPS.

The TV-prior decomposes into a sum where each entry only depends on neighboring points; the uniform blur operator is similarly local. This implies in combination that the posterior can be factorized at granularities defined by the user, and we can therefore apply the local BPS. We stress that the global BPS struggles in high dimensions [24], and thus localization is necessary for it to be a competitive algorithm in these settings. The proximal operator is not available in closed form for the TV-prior, and hence requires evaluation via numerical schemes such as the Douglas-Rachford algorithm introduced in [25] or the Chambolle-Pock algorithm [17]. While these algorithms in general are efficient, they slow down significantly as the precision of the envelope is increased.

We compare the performances of the LBPS, the ZZS, pMALA, MY-ULA, MY-UULA, and SK-ROCK. For both the LBPS and the ZZS we estimated bounds on the prior- and likelihood-gradients, and used these constant bounds to generate computationally cheap events, avoiding any global evaluations of the gradient. For pMALA, we set $\lambda = 2\delta = 0.006$, giving us an acceptance ratio of 67%. For the last three samplers, we chose $\lambda = 0.45$ following the guidance in [27]. The goal is to sample from the posterior distribution when observing a blurred image, see Figure 7. Figure 8 shows the mean squared error (MSE) and the structural similarity index (SSIM) between the mean estimates of the various algorithms and the 'true' mean, as estimated by a long run of an asymptotically exact algorithm. Notably, unlike MY-(U)ULA, pMALA, and SK-ROCK, which require the evaluation of the proximal operator (which is not localizable), the LBPS and ZZS can be sped up using parallelization techniques: the implementation we used applied global rates to avoid recalculating the full posterior gradient after every event, but one may calculate the factor gradients at hardly any extra computational cost if one calculates them in parallel.

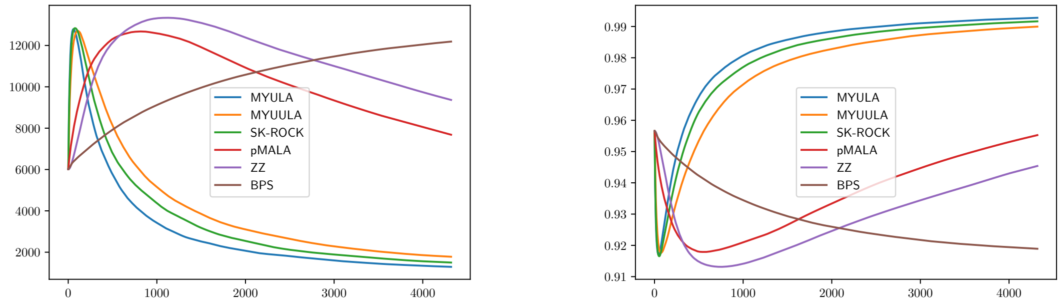

Figure 8: Results from the Image Deblurring example. Left: The MSE of the mean estimates, estimated every 10 seconds. Right: The SSIM of the mean estimates, estimated every 10 seconds.