8. Appendix

This appendix contains additional plots and proofs of the results from Section 2.

Lemma 6. *The divergence from $q(z)$ to $p(z)$ is*

$$KL\left(q(Z)\|p(Z)\right)=\underbrace{KL\left(q(Z|W)\|p(Z)\right)}_{D_0}-I_q[W,Z], \tag{27}$$

where $D_0=\mathbb{E}_{q(W,Z)}\log\left(q(Z|W)/p(Z)\right)$ is conditional divergence and I_q denotes mutual information under q.

Proof. Define the joint distribution $p(w,z)=q(w)p(z)$. Then, the chain-rule of KL-divergence (Cover & Thomas, 2006, Thm. 2.5.3) states that

$$\begin{aligned}KL\left(q(Z,W)\|p(Z,W)\right)=&KL\left(q(W|Z)\|p(W|Z)\right)\\&+KL\left(q(Z)\|p(Z)\right).\end{aligned} \tag{28}$$

The left-hand side simplifies into D_0, and the first term on the right-hand side simplifies into $I_q[W,Z]$. □

Theorem 7. *For fixed values of β and $p(w|z)$, the distribution $q(w)$ that minimizes D_β is*

$$\begin{aligned}q^*(w)&=\exp\big(s(w)-A\big)\\A&=\log\int_w\exp s(w)\\s(w)&=\log p(w)-KL\left(q(Z|w)\|p(Z|w)\right)\\&\quad-\left(\beta^{-1}-1\right)KL\left(q(Z|w)\|p(Z)\right).\end{aligned}$$

Moreover, at q^*, the objective value is $D^*_\beta=-\beta A$.

Proof. First, consider derivatives of D_0 and D_1 with respect to $q(w)$. The first can immediately be seen to be

$$\frac{dD_0}{dq(w)}=KL\left(q(Z|w)\|p(Z)\right).$$

For the second, we can derive

$$\begin{aligned}\frac{dD_1}{dq(w)}=&\frac{d}{dq(w)}\int_{w,z}q(w,z)\log\frac{q(z|w)}{p(w,z)}\\&+\frac{d}{dq(w)}\int_{w,z}q(w)\log q(w)\\=&\int_z q(z|w)\log\frac{q(z|w)}{p(w,z)}+\log q(w)+1\\=&KL\left(q(Z|w)\|p(Z|w)\right)-\log p(w)+\log q(w)+1.\end{aligned}$$

If we create a Lagrangian for D_β with a Lagrange multiplier λ to enforce normalization of $q(w)$, we know that at the optimal $q(w)$ its gradient will be zero. Using the above derivatives, we therefore have that

$$\begin{aligned}0=&(1-\beta)KL\left(q(Z|w)\|p(Z)\right)+\beta KL\left(q(Z|w)\|p(Z|w)\right)\\&-\beta\log p(w)+\beta\log q(w)+\lambda,\end{aligned}$$

Which solved for $q(w)$, this gives

$$\begin{aligned}q(w)\propto\exp\big(&-(1-\beta^{-1})KL\left(q(Z|w)\|p(Z)\right)\\&-KL\left(q(Z|w)\|p(Z|w)\right)+\log p(w)\big),\end{aligned}$$

which establishes the given form for $s(w)$ and A.

Now, to establish the value of D_β at the solution, expand the negative entropy of $q(w)$ to get

$$\begin{aligned}&\beta\int_w q(w)\log q(w)\\&=\beta\int_w q(w)\Big(-(1-\beta^{-1})KL\left(q(Z|w)\|p(Z)\right)\\&\quad-KL\left(q(Z|w)\|p(Z|w)\right)+\log p(w)\Big)-\beta A.\end{aligned} \tag{29}$$

Now, taking the left-hand side and terms in the bottom line, we can recognize that

$$\int_w q(w)\left(\log\frac{p(w)}{q(w)}-KL\left(q(Z|w)\|p(Z|w)\right)\right)=-D_1.$$

Further, if we take the terms from the middle line, we have that

$$-\beta\int_w q(w)(1-\beta^{-1})KL\left(q(Z|w)\|p(Z)\right)=(\beta-1)D_0.$$

Thus, we can re-write Eq. 29 as $-\beta A=(1-\beta)D_0+\beta D_1$, establishing the value of D^*_β. □

Remark 8. In the limit where $\beta\to0$ the divergence bound becomes

$$\lim_{\beta\to0}D^*_\beta=\inf_w KL\left(q(Z|w)\|p(Z)\right).$$

Proof. Use the representation that $\lim_{\beta\to0}D^*_\beta=\lim_{\beta\to0}-\beta A$ is equal to

$$\begin{aligned}&\lim_{\beta\to0}-\beta\log\int_w\exp\Big(\log p(w)-KL\left(q(Z|w)\|p(Z|w)\right)\\&\qquad-\left(\beta^{-1}-1\right)KL\left(q(Z|w)\|p(Z)\right)\Big)\\&=\lim_{\beta\to0}-\beta\log\int_w\exp\Big(-\beta^{-1}KL\left(q(Z|w)\|p(Z)\right)\Big).\end{aligned}$$

The form for D^*_β follows from the fact that $\lim_{\beta\to0}\beta\log\int_w\exp(\beta^{-1}f(w))=\sup_w f(w)$. □

Lemma 9. *If $p(w|z) = r(w)q(z|w)/r_z$ and r_z is a constant, then the solution in Thm. 3 holds with*

$$s(w) = \log r(w) - \log r_z + \mathbb{E}_{q_w(Z)}[\beta^{-1} \log p(z) + (1-\beta^{-1}) \log q(z|w)].$$

Proof. First, without using the particular form for $p(w|z)$, we can write $s(w)$ as

$$\log p(w) - \int_z q(z|w) \log \frac{q(z|w)}{p(z|w)} - \left(\beta^{-1}-1\right) \int_z q(z|w) \log \frac{q(z|w)}{p(z)}$$

Cancelling terms involving $q(z|w)$ in the numerators, this is

$$\log p(w) - \int_z q(z|w) \log \frac{p(z)}{p(z|w)} - \beta^{-1} \int_z q(z|w) \log \frac{q(z|w)}{p(z)}$$

The $\log p(w)$ can be absorbed into the first term to give, after some cancellation that

$$s(w) = \int_z q(z|w) \log p(w|z) - \beta^{-1} KL\left(q(Z|w) \| p(Z)\right).$$

Now, using the assumed form for $p(w|z)$, we can immediately write that $s(w)$ is

$$\int_z q(z|w) \log \frac{r(w)q(z|w)}{r_z} - \beta^{-1} \int_z q(z|w) \log \frac{q(z|w)}{p(z)},$$

equivalent to the form stated. □

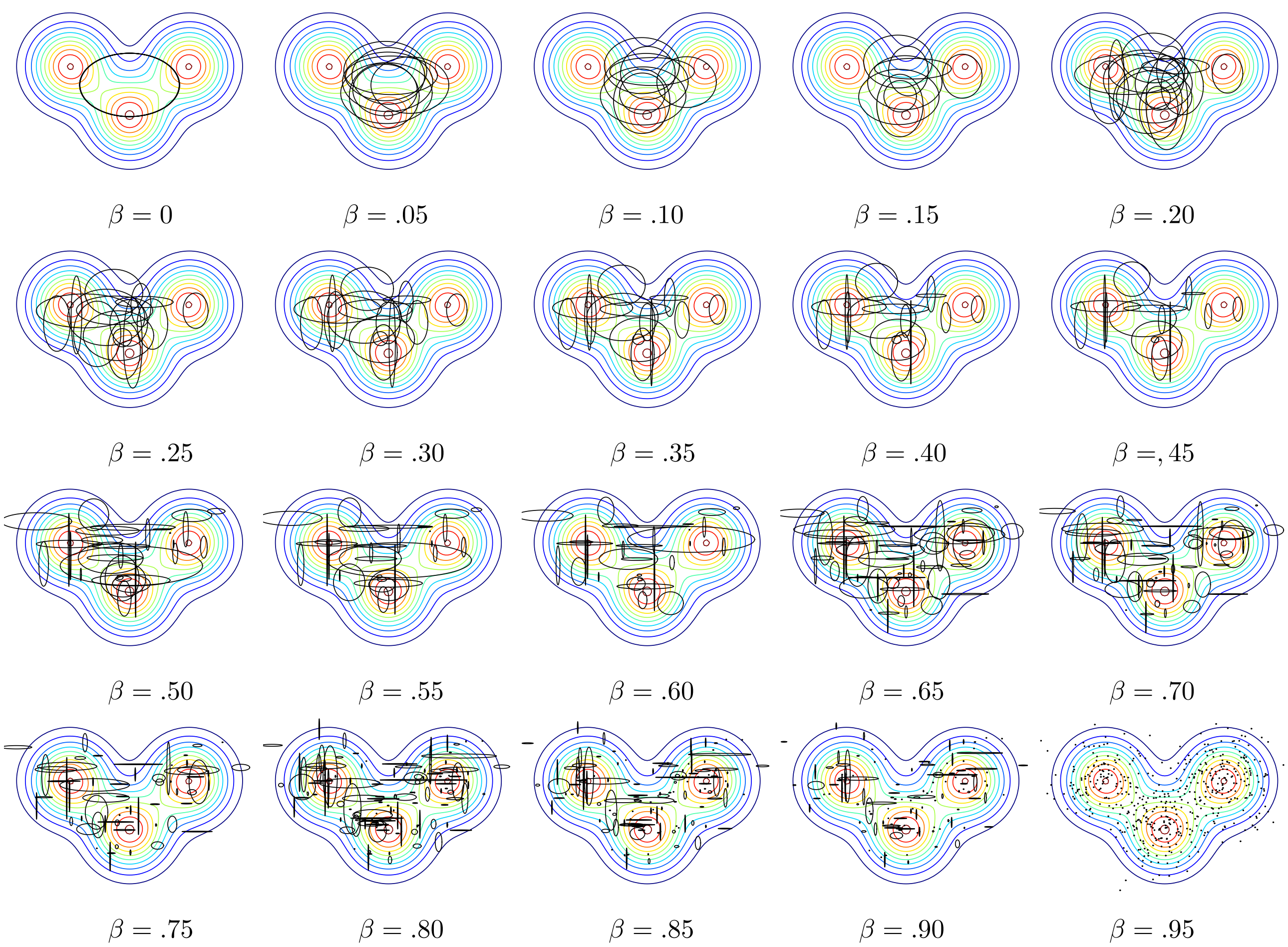

Figure 5. Examples sampling from a two-dimensional mixture of three gaussians after running inference for 5×10^5 iterations. The sampled weights w are pictured as ellipsoids at one standard deviation. Colored contours show the density $p(z)$. To avoid visual clutter, a smaller number (equally spaced) of samples are shown for smaller β.

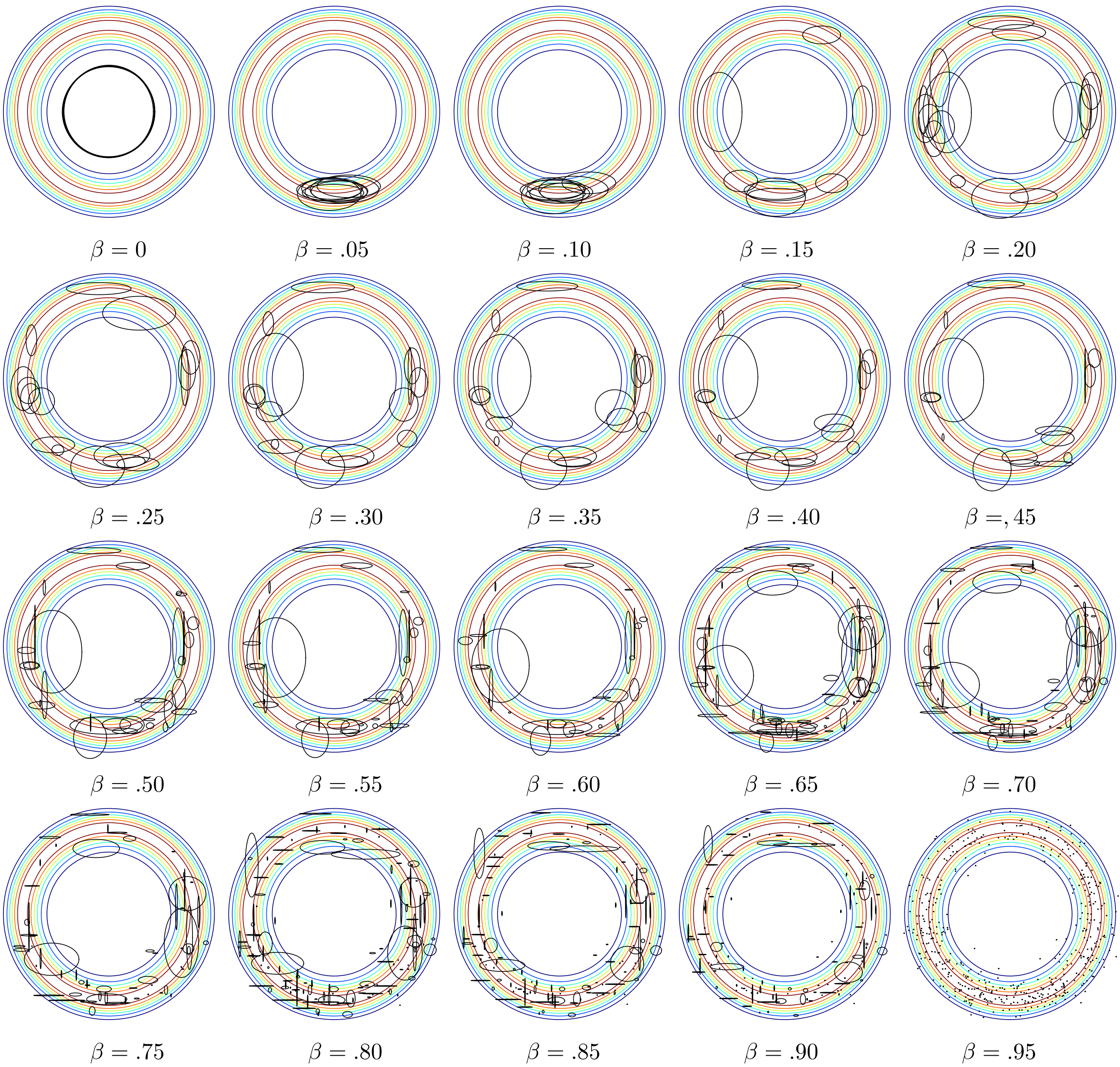

Figure 6. Examples sampling from a two-dimensional "donut" distribution after running inference for 5×10^5 iterations. The sampled weights w are pictured as ellipsoids at one standard deviation. Colored contours show the density $p(z)$. To avoid visual clutter, a smaller number (equally spaced) of samples are shown for smaller β.

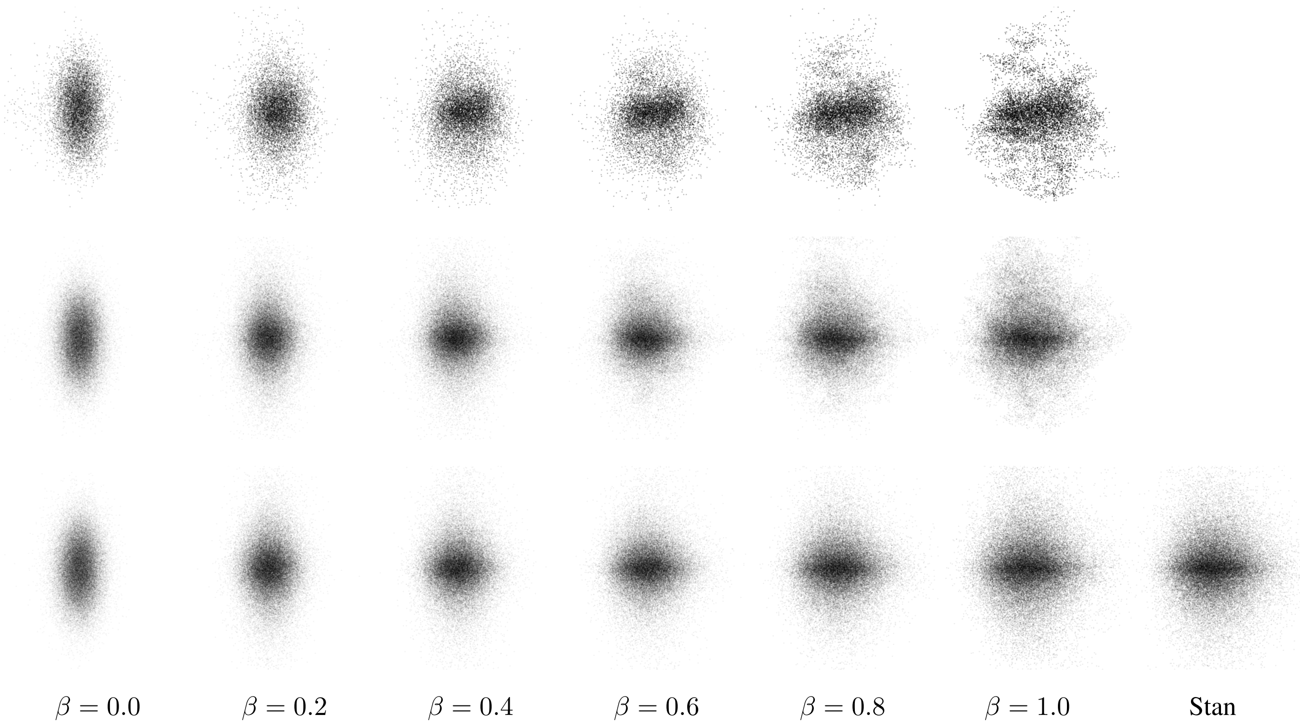

Figure 7. Inference for various values of β on `ionosphere` after 10^4 (top row) 10^5 (middle row) or 10^6 (bottom row) iterations. After each iteration, one sample is drawn from $q_w(Z)$, and plots show the first two principal components (computed on samples from Stan). Each plot show samples resulting from the (constant) step-size ϵ that resulted in the minimum MMD for that β and number of iterations. The same sequence of random numbers is for all inference methods. (More results are in the appendix.)

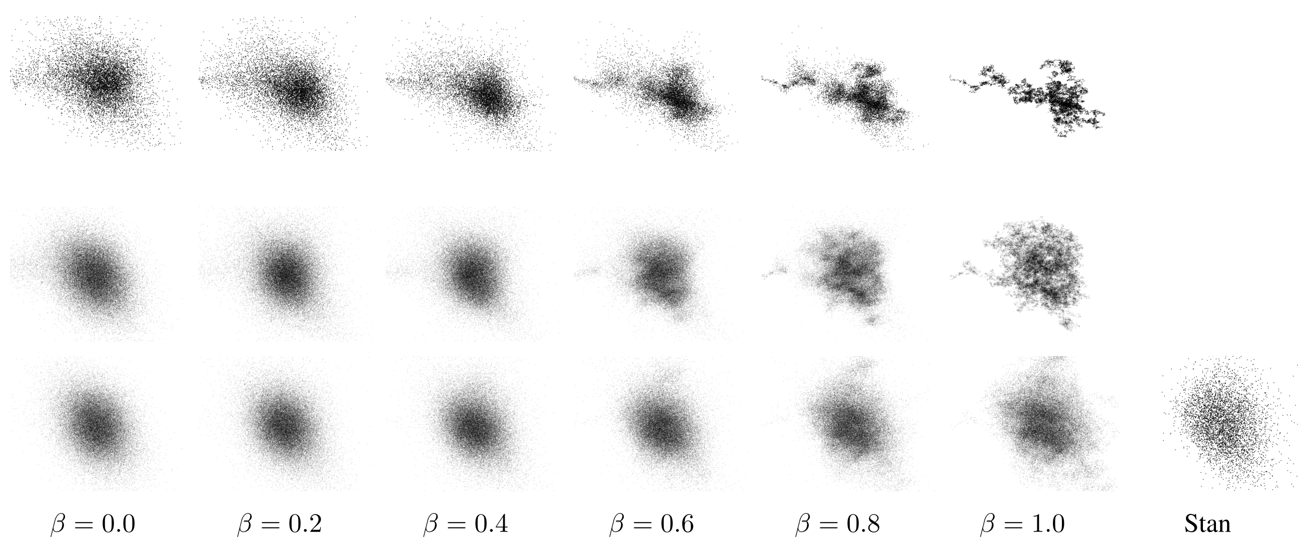

Figure 8. Inference for various values of β on `a1a` after 10^4 (top row) 10^5 (middle row) or 10^6 (bottom row) iterations. In some of these plots, a "tail" is visible, reflecting the path into the high-density region from where $w = 0$ where inference was initialized.

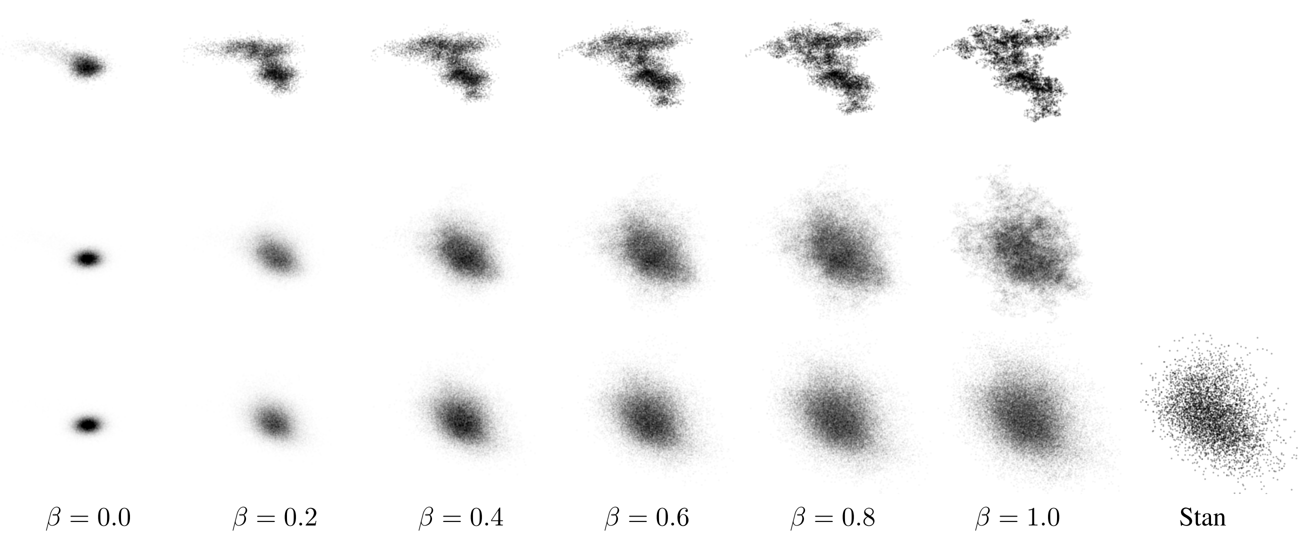

Figure 9. Inference for various values of β on `australian` after 10^4 (top row) 10^5 (middle row) or 10^6 (bottom row) iterations.

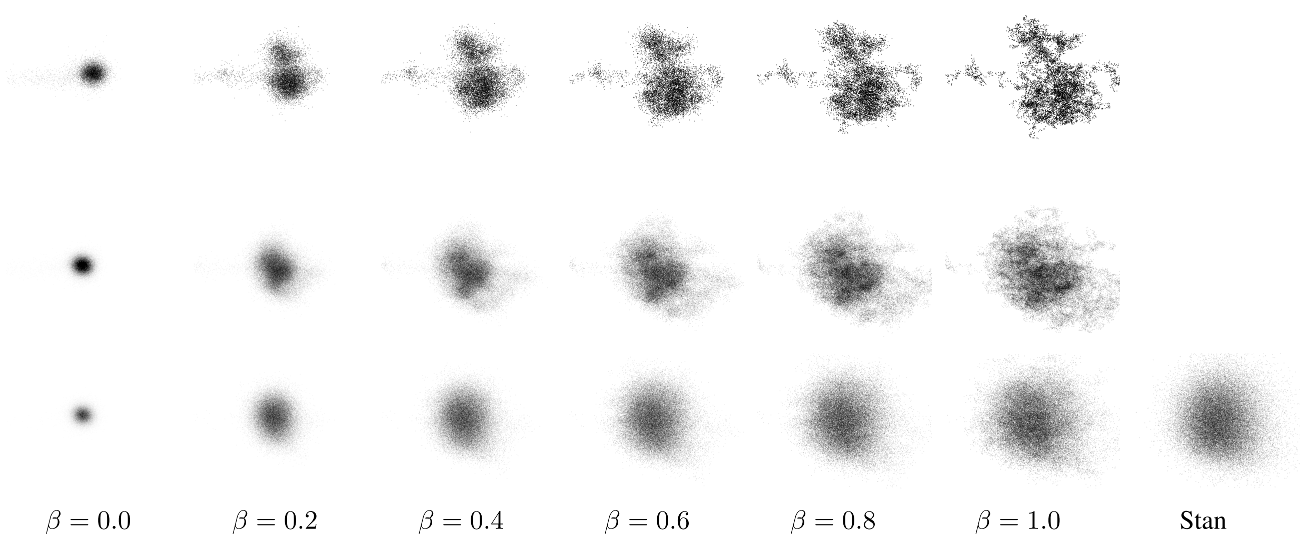

Figure 10. Inference for various values of β on `sonar` after 10^4 (top row) 10^5 (middle row) or 10^6 (bottom row) iterations.