Some sampling methods for pyroclastic flows

Goal Efficiently compute the probability of rare events

Importance sampling Sample from an alternate/biased distribution in order to make rare (yet desired) events occur frequently. Consider

$$
\begin{aligned}
P(\gamma > L) &= \int \mathrm{I}(\gamma(x))p(x)dx && (1)\\
&= \int \mathrm{I}(\gamma(x))\frac{p(x)}{p^*(x)}p^*(x)dx && (2)
\end{aligned}
$$

where I is an indicator function. Or consider the Monte Carlo (MC) approximation

$$
\begin{aligned}
P(\gamma > L) &\approx \frac{1}{N}\sum_{k=1}^{N}\mathrm{I}(\gamma(X_k)) && (3)\\
&= \frac{1}{N}\sum_{k=1}^{N}\mathrm{I}(\gamma(X_k))\frac{p(X_k)}{p^*(X_k)} && (4)
\end{aligned}
$$

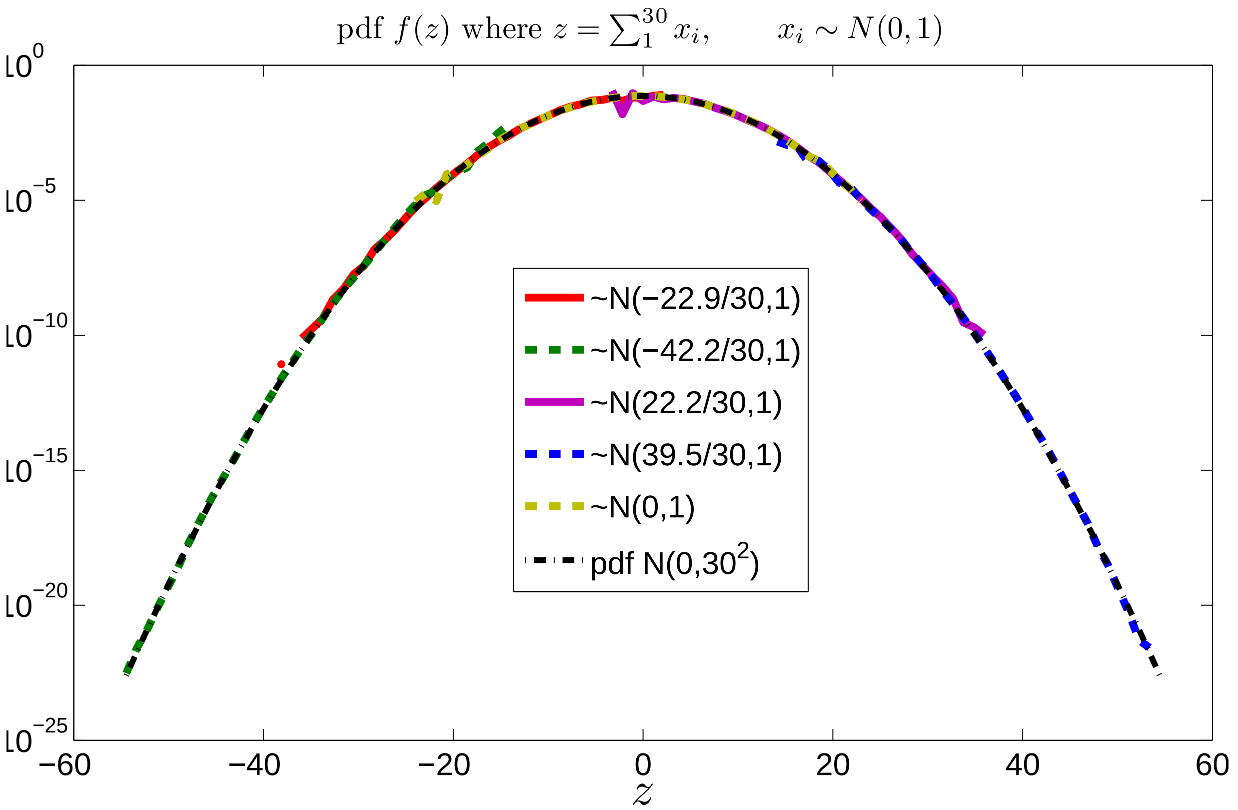
where $X_k \sim p$ in Equation (3) and $X_k \sim p^*$ in Equation (4), but biasing is corrected with the *likelihood ratio*, $lr = p/p^*$.

Example: sum of 30 Gaussian rvs

$$z = x_1 + x_2 + \ldots + x_{30}, \quad x_k \sim N(0,1) \Rightarrow z \sim N(0,30)$$

- Suppose we wish to simulate the pdf of z to small probabilities: proposed biased distribution(s) $x_k \sim N(v,1)$ i.e. $p^* = N(v,1)$
- For targeted sum, say $z = C$, Lagrange multiplier problem shows optimal mean shift of $v = \dfrac{C}{30}$ for each of the 30 rvs.
- Can also choose v, the *biasing* or *tilting parameter*, by numerical methods — Cross-Entropy

Result from simulation



- targeted sums of $\{-20, -40, 20, 40, 0\}$
- biasing parameters chosen here by cross-entropy
- each simulation 10,000 samples

Mini intro to Cross-Entropy (Rubinstein and Kroese)

Goal pick a distribution to sample rare events of interest
Subgoal pick “good” biasing parameters from a family of distributions
our example: family – Gaussian, biasing parameters – mean.
Idea Given some “physical” distribution, g, pick a candidate parameterized distribution, $h(\cdot; \mathbf{v})$. Choose a specific parameter(s) that minimizes the Kullback-Leibler distance

$$\mathcal{D}(g, h) = \mathbb{E}_g \ln \frac{g(X)}{h(X; \mathbf{v})} = \int g \ln(g) dx - \int g \ln(h(x; \mathbf{v})) dx. \tag{5}$$

Since g is given, minimizing Equation (5) is equivalent to

$$\max_{v} \int g(x) \ln(h(x; \mathbf{v})) dx \tag{6}$$

Keep in mind Want distributions to be “close” for X such that $\gamma(X) > L$
-Rewrite optimization problem (Equation (6)) to reflect this goal.

CE continued

$$\max_v \int \mathrm{I}(\gamma(x) > L) g(x) \ln(h(x;\mathbf{v}))dx = \max_v \mathbb{E}_g[\mathrm{I}(\gamma(X) > L) \ln(h(X;\mathbf{v}))]$$

<u>Trick</u> Use importance sampling again, rewrite as

$$\max_v \mathbb{E}_{h(\cdot;\mathrm{w})}[\mathrm{I}(\gamma(X) > L) \frac{g(X)}{h(X;\mathbf{w})} \ln(h(X;(\mathbf{v})))] \tag{7}$$

Optimal **v** can be approximated by solving

$$\frac{1}{N}\sum_{k=1}^{N} \mathrm{I}(\gamma(X_k) > L) \frac{g(X_k)}{h(X_k;\mathbf{w})} \nabla \ln(h(X_k;\mathbf{v})) = \mathbf{0} \tag{8}$$

<u>Issue</u> still too hard if $h(\cdot;\mathbf{w})$ doesn't give us many "hits", i.e., $\gamma(X_k) > L$.

<u>To overcome problem</u>

1) Guess well, or

2) Choose sequence of targets $L_1 < L_2 < \ldots < L_{j-1} < L_j = L$ such that $P(\gamma > L_1)$ is not too small. Find $\mathbf{v}_1$ with above method (might not need IS in first step). For second step, i.e., want to find $\mathbf{v}_2$ based on target of L_2, use $\mathbf{w} = \mathbf{v}_1$. Repeat as necessary.

Back to sum of Gaussian example

Recall problem:

$$z = \gamma(\mathbf{x}) = x_1 + x_2 + \ldots + x_{30}, \quad x_k \sim N(0,1) \Rightarrow z \sim N(0,30)$$

In our case

$$g(\mathbf{x}) = \frac{1}{(2\pi)^{15}} \exp(-\mathbf{x}\mathbf{x}^T/2), \quad h(\mathbf{x};\mathbf{v}) = \frac{1}{(2\pi)^{15}} \exp(-(\mathbf{x}-\mathbf{v})(\mathbf{x}-\mathbf{v})^T/2)$$

Find **v** such that for each j

$$\sum_{i=1}^{N} \mathrm{I}(\gamma(\mathbf{X}_i > L))\frac{f(\mathbf{X}_\mathrm{i})}{h(\mathbf{X}_\mathrm{i};\mathbf{w})}(X_{i,j} - v_j) = 0.$$

Thus we have

$$v_j = \frac{\sum_{i=1}^{N} \mathrm{I}(\gamma(\mathbf{X}_i) > L) X_{i,j}\; f(\mathbf{X}_\mathrm{i})/h(\mathbf{X}_\mathrm{i};\mathbf{w})}{\sum_{i=1}^{N} \mathrm{I}(\gamma(\mathbf{X}_i) > L) f(\mathbf{X}_\mathrm{i})/h(\mathbf{X}_\mathrm{i};\mathbf{w})} \tag{9}$$

Slightly more interesting CE example with "guessing"

Consider following random position problem:

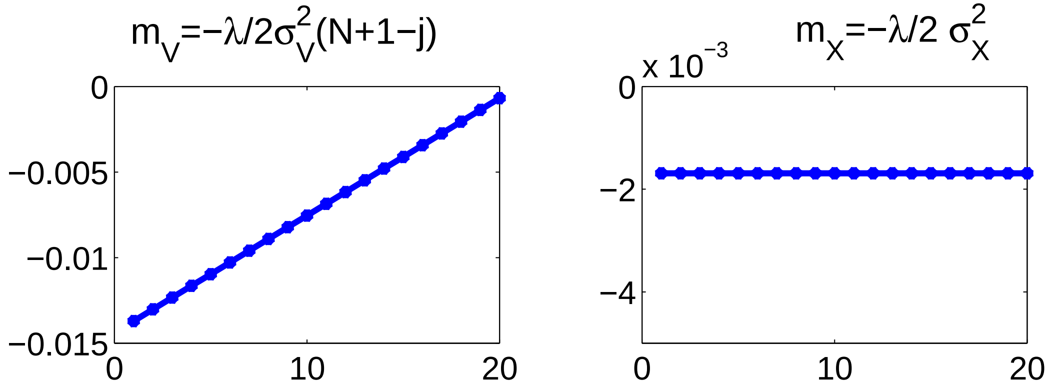
$$V_{k+1} = V_k + \delta V_k \quad \text{and} \quad X_{k+1} = X_k + \Delta T V_k + \delta X_k, \quad X_0 = V_0 = 0, \quad k = 1, \ldots, 20$$

$$\delta V \sim N(0, \sigma_V^2) \quad \text{and} \quad \delta X \sim N(0, \sigma_X^2)$$

- Want to bias mean of δX_k and δV_k in an optimal way as to sample $X_N > X_{target}$ ($N = 20$) and hence find the probability of that event.

- Can write as constrained optimization problem and solve analytically.

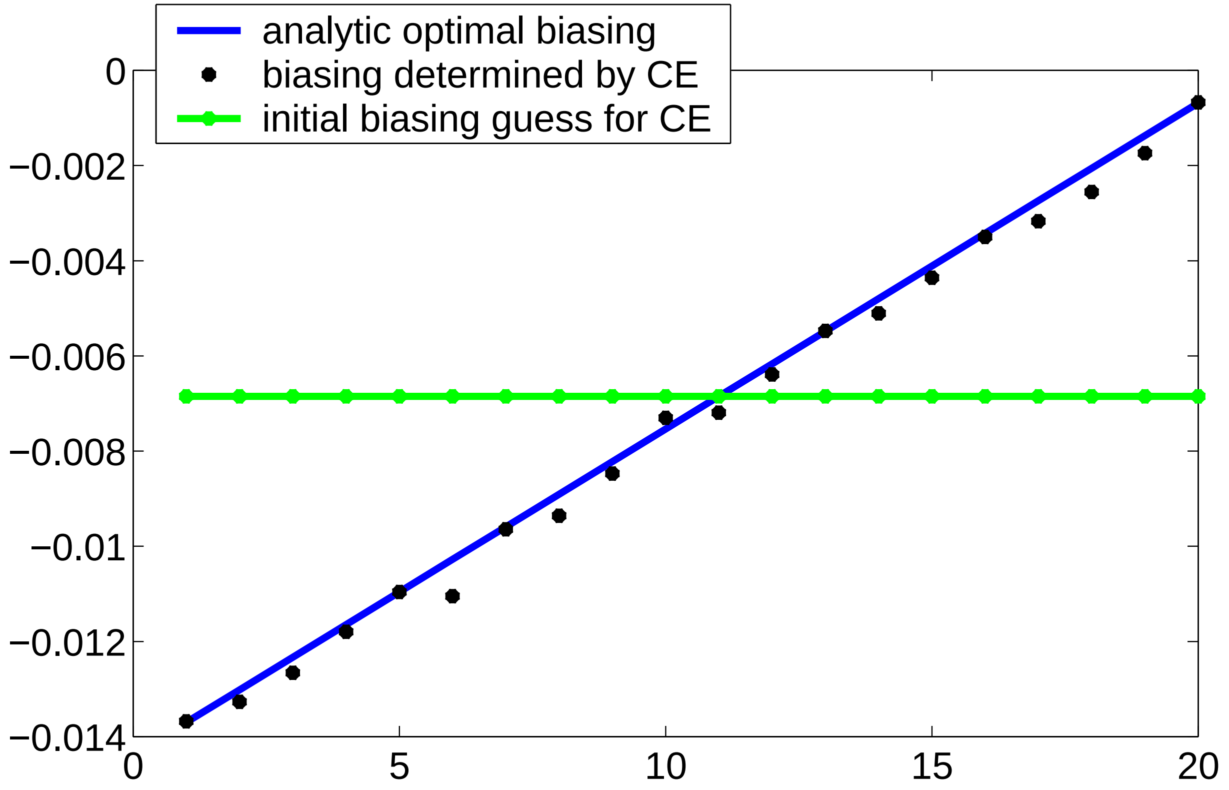
- Solution: constant for biasing δX, linear biasing for δV



$\lambda = -2X_{target}/(N(\sigma_X^2 + \Delta T^2 \sigma_Z^2 (N+1)(2N+1)/6))$

Focus on velocity biasing

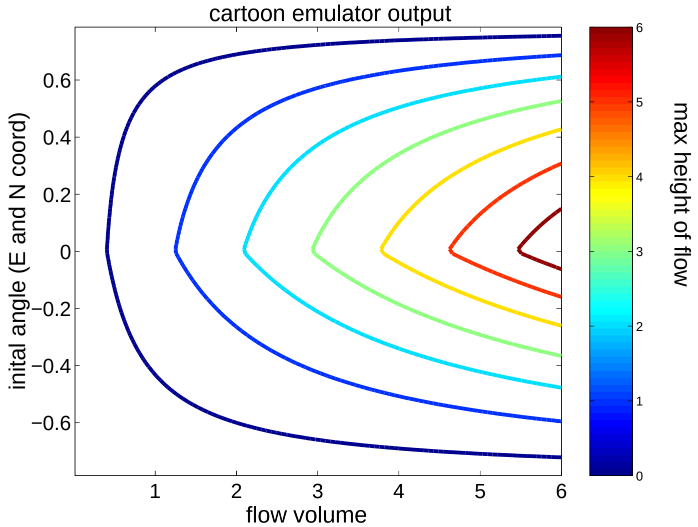
- Start with a biased velocity distribution with constant mean
- Update with Equation (9), but now $\mathrm{I}(X_N + a > X_{target} > X_N - a)$



Basic P-flow sampling problem

1. Choose a location where we'd like to know $P(h_{max} > h_{crit})$ over some time period, say 100 years.

2. Construct an emulator over a range of initial volumes and coordinates that cover a "reasonable" h_{max} output, say $[\varepsilon, 2h_{crit}]$.



3. Find decent biasing distributions by using IS and CE.

4. "Unlimited" sample size if we evaluate emulator in MC schemes.

Problems/questions

1. How to sample from Robert's volume process?

2. How to "train" CE algorithm if $P(h_{max} > \varepsilon)$ is really small? That is, probability of no flow at target is high?

3. How can we reflect emulator uncertainty as uncertainty in probability estimates? Uncertainty in volume distribution?

4. Does it make sense to evaluate with the mean of the emulator output? Or *really* sample the GASP and include randomness in the emulator max height evaluations?

5. Any suggestions for reading or ideas for toy problems?