Chapter 22

Large Deviations for Small-Noise Stochastic Differential Equations

This lecture is at once the end of our main consideration of diffusions and stochastic calculus, and a first taste of large deviations theory. Here we study the divergence between the trajectories produced by an ordinary differential equation, and the trajectories of the same system perturbed by a small amount of white noise.

Section 22.1 establishes that, in the small noise limit, the SDE's trajectories converge in probability on the ODE's trajectory. This uses Feller-process convergence.

Section 22.2 upper bounds the rate at which the probability of large deviations goes to zero as the noise vanishes. The methods are elementary, but illustrate deeper themes to which we will recur once we have the tools of ergodic and information theory.

In this chapter, we will use the results we have already obtained about SDEs to give a rough estimate of a basic problem, frequently arising in practice[1] namely taking a system governed by an ordinary differential equation and seeing how much effect injecting a small amount of white noise has. More exactly, we will put an upper bound on the probability that the perturbed trajectory goes very far from the unperturbed trajectory, and see the rate at which this probability goes to zero as the amplitude ϵ of the noise shrinks; this will be

[1] For applications in statistical physics and chemistry, see Keizer (1987). For applications in signal processing and systems theory, see Kushner (1984). For applications in nonparametric regression and estimation, and also radio engineering (!) see Ibragimov and Has'minskii (1979/1981). The last book is especially recommended for those who care about the connections between stochastic process theory and statistical inference, but unfortunately expounding the results, or even just the problems, would require a too-long detour through asymptotic statistical theory.

$O(e^{-C\epsilon^2})$. This will be our first illustration of a large deviations calculation. It will be crude, but it will also introduce some themes to which we will return (inshallah!) at greater length towards the end of the course. Then we will see that the major improvement of the more refined tools is to give a lower bound to match the upper bound we will calculate now, and see that we at least got the logarithmic rate right.

I should say before going any further that this example is shamelessly ripped off from Freidlin and Wentzell (1998, ch. 3, sec. 1, pp. 70–71), which is *the* book on the subject of large deviations for continuous-time processes.

22.1 Convergence in Probability of SDEs to ODEs

To begin with, consider an unperturbed ordinary differential equation:

$$\frac{d}{dt}x(t) = a(x(t)) \tag{22.1}$$

$$x(0) = x_0 \in \mathbb{R}^d \tag{22.2}$$

Assume that a is uniformly Lipschitz-continuous (as in the existence and uniqueness theorem for ODEs, and more to the point for SDEs). Then, for the given, non-random initial condition, there exists a unique continuous function x which solves the ODE.

Now, for $\epsilon > 0$, consider the SDE

$$dX_\epsilon = a(X_\epsilon)dt + \epsilon dW \tag{22.3}$$

where W is a standard d-dimensional Wiener process, with non-random initial condition $X_\epsilon(0) = x_0$. Theorem 216 clearly applies, and consequently so does Theorem 220, meaning X_ϵ is a Feller diffusion with generator $G_\epsilon f(x) = a_i(x)\partial_i f'(x) + \frac{\epsilon^2}{2}\nabla^2 f(x)$.

Write X_0 for the deterministic solution of the ODE.

Our first assertion is that $X_\epsilon \stackrel{d}{\to} X_0$ as $\epsilon \to 0$. Notice that X_0 is a Feller process[2], whose generator is $G_0 = a_i(x)\partial_i$. We can apply Theorem 170 on convergence of Feller processes. Take the class of functions with bounded second derivatives. This is clearly a core for G_0, and for every G_ϵ. For every function f in this class,

$$\|G_\epsilon f - G_0 f\|_\infty = \left\| a_i\partial_i f(x) + \frac{\epsilon^2}{2}\nabla^2 f(x) - a_i\partial_i f(x) \right\|_\infty \tag{22.4}$$

$$= \frac{\epsilon^2}{2}\left\|\nabla^2 f(x)\right\|_\infty \tag{22.5}$$

which goes to zero as $\epsilon \to 0$. But this is condition (i) of the convergence theorem, which is equivalent to condition (iv), that convergence in distribution of the

[2]You can amuse yourself by showing this. Remember that $X_y(t) \stackrel{d}{\to} X_x(t)$ is equivalent to $\mathbf{E}[f(X_t)|X_0 = y] \to \mathbf{E}[f(X_t)|X_0 = x]$ for all bounded *continuous* f, and the solution of an ODE depends continuously on its initial condition.

initial condition implies convergence in distribution of the whole trajectory. Since the initial condition is the same non-random point x_0 for all ϵ, we have $X_\epsilon \xrightarrow{d} X_0$ as $\epsilon \to 0$. In fact, since X_0 is non-random, we have that $X_\epsilon \xrightarrow{P} X_0$. That last assertion really needs some consideration of metrics on the space of continuous random functions to make sense (see Appendix A2 of Kallenberg), but once that's done, the upshot is

Theorem 253 *Let $\Delta_\epsilon(t) = |X_\epsilon(t) - X_0(t)|$. For every $T > 0$, $\delta > 0$,*

$$\lim_{\epsilon \to 0} \mathbb{P}\left(\sup_{0 \leq t \leq T} \Delta_\epsilon(t) > \delta\right) = 0 \tag{22.6}$$

Or, using the maximum-process notation, for every $T > 0$,

$$\Delta(T)^* \xrightarrow{P} 0 \tag{22.7}$$

Proof: See above. □

This is a version of the weak law of large numbers, and nice enough in its own way. One crucial limitation, however, is that it tells us nothing about the *rate* of convergence. That is, it leaves us clueless about how big the noise can be, while still leaving us in the small-noise limit. If the rate of convergence were, say, $O(\epsilon^{1/100})$, then this would not be very useful. (In fact, if the convergence were that slow, we should be really suspicious of numerical solutions of the *unperturbed* ODE.)

22.2 Rate of Convergence; Probability of Large Deviations

Large deviations theory is essentially a study of rates of convergence in probabilistic limit theorems. Here, we will estimate the rate of convergence: our methods will be crude, but it will turn out that even more refined estimates won't change the rate, at least not by more than log factors.

Let's go back to the difference between the perturbed and unperturbed trajectories, going through our now-familiar procedure.

$$X_\epsilon(t) - X_0(t) = \int_0^t [a(X_\epsilon(s)) - a(X_0(s))]\, ds + \epsilon W(t) \tag{22.8}$$

$$\Delta_\epsilon(t) \leq \int_0^t |a(X_\epsilon(s)) - a(X_0(s))|\, ds + \epsilon |W(t)| \tag{22.9}$$

$$\leq K_a \int_0^t \Delta_\epsilon(s) ds + \epsilon |W(t)| \tag{22.10}$$

$$\Delta_\epsilon^*(T) \leq \epsilon W^*(T) + K_a \int_0^t \Delta_\epsilon^*(s) ds \tag{22.11}$$

Applying Gronwall's Inequality (Lemma 214),

$$\Delta_\epsilon^*(T) \leq \epsilon W^*(T) e^{K_a T} \tag{22.12}$$

The only random component on the RHS is the supremum of the Wiener process, so we're in business, at least once we take on two standard results, one about the Wiener process itself, the other just about multivariate Gaussians.

Lemma 254 *For a standard Wiener process,* $\mathbb{P}(W^*(t) > a) = 2\mathbb{P}(|W(t)| > a)$.

PROOF: Proposition 13.13 (pp. 256–257) in Kallenberg. □

Lemma 255 *If Z is a d-dimensional standard Gaussian (i.e., mean 0 and covariance matrix I), then*

$$\mathbb{P}(|Z| > z) \leq \frac{2z^{d-2}e^{-z^2/2}}{2^{d/2}\Gamma(d/2)} \tag{22.13}$$

for sufficiently large z.

PROOF: Each component of Z, $Z_i \sim \mathcal{N}(0,1)$. So $|Z| = \sqrt{\sum_{i=1}^{d} Z_i^2}$ has the density function (see, e.g., (Cramér, 1945, sec. 18.1, p. 236))

$$f(z) = \frac{2}{2^{d/2}\sigma^d\Gamma(d/2)} z^{d-1} e^{-\frac{z^2}{2\sigma^2}}$$

This is the d-dimensional Maxwell-Boltzmann distribution, sometimes called the χ-distribution, because $|Z|^2$ is χ^2-distributed with d degrees of freedom. Notice that $\mathbb{P}(|Z| \geq z) = \mathbb{P}\left(|Z|^2 \geq z^2\right)$, so we will be able to solve this problem in terms of the χ^2 distribution. Specifically, $\mathbb{P}\left(|Z|^2 \geq z^2\right) = \Gamma(d/2, z^2/2)/\Gamma(d/2)$, where $\Gamma(r,a)$ is the upper incomplete gamma function. For said function, for every r, $\Gamma(r,a) \leq a^{r-1}e^{-a}$ for sufficiently large a (Abramowitz and Stegun, 1964, Eq. 6.5.32, p. 263). Hence (for sufficiently large z)

$$\begin{aligned}
\mathbb{P}(|Z| \geq z) &= \mathbb{P}\left(|Z|^2 \geq z^2\right) && (22.14)\\
&= \frac{\Gamma(d/2, z^2/2)}{\Gamma(d/2)} && (22.15)\\
&\leq \frac{\left(z^2\right)^{d/2-1} 2^{1-d/2} e^{-z^2/2}}{\Gamma(d/2)} && (22.16)\\
&= \frac{2z^{d-2}e^{-z^2/2}}{2^{d/2}\Gamma(d/2)} && (22.17)
\end{aligned}$$

□

Theorem 256 *In the limit as $\epsilon \to 0$, for every $\delta > 0$, $T > 0$,*

$$\log \mathbb{P}(\Delta_\epsilon^*(T) > \delta) \leq O(\epsilon^{-2}) \tag{22.18}$$

PROOF: Start by directly estimating the probability of the deviation, using preceding lemmas.

$$\begin{aligned}
\mathbb{P}\left(\Delta_\epsilon^*(T) > \delta\right) &\leq \mathbb{P}\left(|W|^*(T) > \frac{\delta e^{-K_a T}}{\epsilon}\right) && (22.19)\\
&= 2\mathbb{P}\left(|W(T)| > \frac{\delta e^{-K_a T}}{\epsilon}\right) && (22.20)\\
&\leq \frac{4}{2^{d/2}\Gamma(d/2)}\left(\frac{\delta^2 e^{-2K_a T}}{\epsilon^2}\right)^{d/2-1} e^{-\frac{\delta^2 e^{-2K_a T}}{2\epsilon^2}} && (22.21)
\end{aligned}$$

if ϵ is sufficiently small, so that ϵ^{-1} is sufficiently large to apply Lemma 255. Now take the log and multiply through by ϵ^2:

$$\begin{aligned}
&\epsilon^2 \log \mathbb{P}\left(\Delta_\epsilon^*(T) > \delta\right) && (22.22)\\
&\leq \epsilon^2 \log \frac{4}{2^{d/2}\Gamma(d/2)} + \epsilon^2\left(\frac{d}{2} - 1\right)\left[\log \delta^2 e^{-2K_a T} - 2\log\epsilon\right] - \delta^2 e^{-2K_a T}
\end{aligned}$$

$$\lim_{\epsilon \downarrow 0} \epsilon^2 \log \mathbb{P}\left(\Delta_\epsilon^*(T) > \delta\right) \leq -\delta^2 e^{-2K_a T} \quad (22.23)$$

since $\epsilon^2 \log \epsilon \to 0$, and the conclusion follows. □

Notice several points.

1. Here ϵ gauges the size of the noise, and we take a small noise limit. In many forms of large deviations theory, we are concerned with large-sample ($N \to \infty$) or long-time ($T \to \infty$) limits. In every case, we will identify some asymptotic parameter, and obtain limits on the asymptotic probabilities. There are deviations inequalities which hold *non*-asymptotically, but they have a different flavor, and require different machinery. (Some people are made uncomfortable by an ϵ^2 rate, and prefer to write the SDE $dX = a(X)dt + \sqrt{\epsilon}dW$ so as to avoid it. I don't get this.)

2. The magnitude of the deviation δ does not change as the noise becomes small. This is basically what makes this a *large* deviations result. There is also a theory of *moderate* deviations, which with any luck we'll be able to at least touch on.

3. We only have an upper bound. This is enough to let us know that the probability of large deviations becomes exponentially small. But we might be wrong about the rate — it could be even faster than we've estimated. In this case, however, it'll turn out that we've got at least the order of magnitude correct.

4. We also don't have a *lower* bound on the probability, which is something that would be *very* useful in doing reliability analyses. It will turn out that, under many circumstances, one can obtain a lower bound on the probability of large deviations, which has the *same* asymptotic dependence on ϵ as the upper bound.

5. Suppose we're right about the rate (which, it will turn out, we are), and it holds both from above and below. It would be nice to be able to say something like

$$\mathbb{P}\left(\Delta_\epsilon^*(T) > \delta\right) \to C_1(\delta, T)e^{-C_2(\delta,T)\epsilon^{-2}} \tag{22.24}$$

rather than

$$\epsilon^2 \log \mathbb{P}\left(\Delta_\epsilon^*(T) > \delta\right) \to -C_2(\delta, T) \tag{22.25}$$

The difficulty with making an assertion like 22.24 is that the large deviation *probability* actually converges on *any* function which goes to asymptotically to zero! So, to extract the actual rate of dependence, we need to get a result like 22.25.

More generally, one consequence of Theorem 256 is that SDE trajectories which are far from the trajectory of the ODE have exponentially small probabilities. The vast majority of the probability will be concentrated around the unperturbed trajectory. Reasonable sample-path functionals can therefore be well-approximated by averaging their value over some small (δ) neighborhood of the unperturbed trajectory. This should sound very similar to Laplace's method for the evaluate of asymptotic integrals in Euclidean space, and in fact one of the key parts of large deviations theory is an extension of Laplace's method to infinite-dimensional function spaces.

In addition to this mathematical content, there is also a close connection to the principle of least action in physics. In classical mechanics, the system follows the trajectory of least action, the "action" of a trajectory being the integral of the kinetic minus potential energy along that path. In quantum mechanics, this is no longer an axiom but a *consequence* of the dynamics: the action-minimizing trajectory is the most probable one, and large deviations from it have exponentially small probability. Similarly, the theory of large deviations can be used to establish quite general *stochastic* principles of least action for Markovian systems.[3]

[3]For a fuller discussion, see Eyink (1996),Freidlin and Wentzell (1998, ch. 3).