Definition 1.67: From Markov kernels to linear operators

Let $p(s,\mathbf{x},t,\mathrm{d}\mathbf{y})$ be a Markov kernel that satisfies the Chapman-Kolmogorov equation and initial condition:

$$\forall s<\tau<t,\ p(s,\mathbf{x},t,\mathrm{d}\mathbf{y})=\int p(s,\mathbf{x},\tau,\mathrm{d}\mathbf{z})p(\tau,\mathbf{z},t,\mathrm{d}\mathbf{y}), \tag{1.13}$$

$$p(s,\mathbf{x},s,\mathrm{d}\mathbf{y})=\delta_{\mathbf{x}}(\mathrm{d}\mathbf{y}). \tag{1.14}$$

A Markov kernel induces two bounded (in fact, contractive) and positive linear opeators (for any $s\le t$):

$$\mathsf{T}_{s,t}:\mathcal{B}(\Sigma)\to\mathcal{B}(\Sigma),\ \ \mathsf{T}_{s,t}f(\mathbf{x})=(\mathsf{T}_{s,t}f)(\mathbf{x}):=\int p(s,\mathbf{x},t,\mathrm{d}\mathbf{y})f(\mathbf{y}), \tag{1.15}$$

$$\mathsf{T}^*_{s,t}:\mathcal{M}(\Sigma)\to\mathcal{M}(\Sigma),\ \ \mu\mathsf{T}^*_{s,t}(\mathrm{d}\mathbf{y})=(\mu\mathsf{T}^*_{s,t})(\mathrm{d}\mathbf{y}):=\int p(s,\mathbf{x},t,\mathrm{d}\mathbf{y})\mu(\mathrm{d}\mathbf{x}).$$

It follows from the Chapman-Kolmogorov equation (1.13) and the initial condition (1.14) that

$$\mathsf{T}_{s,t}=\mathsf{T}_{s,\tau}\mathsf{T}_{\tau,t},\qquad \mathsf{T}^*_{s,t}=\mathsf{T}^*_{s,\tau}\mathsf{T}^*_{\tau,t},\qquad \mathsf{T}_{t,t}=\mathsf{T}^*_{s,s}=\mathrm{Id}.$$

As the notation suggests, the following adjoint relation follows from Fubini's theorem:

$$\langle\mathsf{T}_{s,t}f;\mu\rangle=\left\langle f;\mu\mathsf{T}^*_{s,t}\right\rangle.$$

We call the following "limit" (when exists) the generators of the Markov kernel $p(s,\mathbf{x},t,\mathrm{d}\mathbf{y})$:

$$\mathsf{L}_{t+}:=\lim_{h\downarrow0}\tfrac{\mathsf{T}_{t,t+h}-\mathrm{Id}}{h},\qquad \mathsf{L}^*_{t+}:=\lim_{h\downarrow0}\tfrac{\mathsf{T}^*_{t,t+h}-\mathrm{Id}}{h}, \tag{1.16}$$

$$\mathsf{L}_{s-}:=\lim_{h\downarrow0}\tfrac{\mathsf{T}_{s-h,s}-\mathrm{Id}}{h},\qquad \mathsf{L}^*_{s-}:=\lim_{h\downarrow0}\tfrac{\mathsf{T}^*_{s-h,s}-\mathrm{Id}}{h}. \tag{1.17}$$

(These are clearly linear operators on their domain.) There are two standard topologies that would make sense the above limits:

- uniform (operator) convergence, where the convergence is uniform over a "ball" of functions f or measures μ. However, this notion is usually too strong to be useful for Markov processes.
- pointwise convergence, where the convergence is pointwise for any fixed function f or measure μ. By default this will be what we use below. We have some additional choices on the topology of the range:
  - strong/weak$(-*)$ convergence, where convergence is w.r.t. some norm topology (e.g., the sup norm restricted to a subspace such as $\mathcal{C}^2_c$) or weak$(-*)$ topology induced by a (pre)dual space.
  - pointwise convergence, where convergence is again pointwise for each fixed $\mathbf{x}\in\mathbb{X}$ or $A\in\Sigma$.

Under the PP (pointwise-pointwise) topology, we have again the nice adjoint relations:

$$\langle\mathsf{L}_{t+}f;\mu\rangle=\left\langle f;\mu\mathsf{L}^*_{t+}\right\rangle,\qquad \langle\mathsf{L}_{s-}f;\mu\rangle=\left\langle f;\mu\mathsf{L}^*_{s-}\right\rangle,$$

where we have pushed the limit inside the pairing $\langle\cdot;\cdot\rangle$.

Theorem 1.68: Diffusion generator of a Markov process (e.g., Skorokhod 1996, p. 161)

Let $A(t,\mathbf{x})\succeq\mathbf{0}$. The generator

$$\mathsf{L}_{t+}f(\mathbf{x})=\tfrac{1}{2}A(t,\mathbf{x})\cdot\nabla^2f(\mathbf{x})+\mathbf{b}(t,\mathbf{x})\cdot\nabla f(\mathbf{x}) \tag{1.18}$$

for all $f\in\mathcal{C}^2_b$ iff the following holds for all t and $\mathbf{x}$:

$$\forall\ \textit{compact nhood } K(\mathbf{x}),\qquad \lim_{h\downarrow0}\tfrac{1}{h}\int_{\mathbb{X}\setminus K(\mathbf{x})}p(t,\mathbf{x},t+h,\mathrm{d}\mathbf{y})=0, \tag{1.19}$$

$$\exists \text{ and hence } \forall \text{ compact nhood } K(\mathbf{x}), \qquad \lim_{h\downarrow 0}\tfrac{1}{h}\int_{K(\mathbf{x})}(\mathbf{y}-\mathbf{x})\cdot p(t,\mathbf{x},t+h,\mathrm{d}\mathbf{y}) = \mathbf{b}(t,\mathbf{x}), \tag{1.20}$$

$$\exists \text{ and hence } \forall \text{ compact nhood } K(\mathbf{x}), \qquad \lim_{h\downarrow 0}\tfrac{1}{h}\int_{K(\mathbf{x})}(\mathbf{y}-\mathbf{x})\otimes(\mathbf{y}-\mathbf{x})\cdot p(t,\mathbf{x},t+h,\mathrm{d}\mathbf{y}) = A(t,\mathbf{x}). \tag{1.21}$$

[We remind that for now this is under the PP (pointwise-pointwise) topology.]

Proof: $\Leftarrow$: Let $f\in\mathcal{C}_b^2$. We apply (1.19) and Taylor expansion on f:

$$\begin{aligned}\tfrac{\mathsf{T}_{t,t+h}-\mathrm{Id}}{h}f(\mathbf{x}) &= \tfrac{1}{h}\int[f(\mathbf{y})-f(\mathbf{x})]\cdot p(t,\mathbf{x},t+h,\mathrm{d}\mathbf{y}) \\ &= \tfrac{1}{h}\int_{K(\mathbf{x})}[f(\mathbf{y})-f(\mathbf{x})]\cdot p(t,\mathbf{x},t+h,\mathrm{d}\mathbf{y}) + o_h(1) \\ &= \tfrac{1}{h}\int_{K(\mathbf{x})}[(\mathbf{y}-\mathbf{x})\cdot\nabla f(\mathbf{x}) + \tfrac{1}{2}(\mathbf{y}-\mathbf{x})\otimes(\mathbf{y}-\mathbf{x})\cdot(\nabla^2 f(\mathbf{x})+o_K(1))]\cdot p(t,\mathbf{x},t+h,\mathrm{d}\mathbf{y}) + o_h(1).\end{aligned} \tag{1.22}$$

Letting $h\downarrow 0$, applying (1.20)-(1.21), and then letting $K\downarrow\mathbf{x}$ proves (1.18).

$\Rightarrow$: We apply the idea of *localization*. Let $f(\mathbf{y})$ be a bump function around $\mathbf{x}$ such that $f(\mathbf{x})=\nabla f(\mathbf{x})=\nabla^2 f(\mathbf{x})=0$, e.g., $f(\mathbf{y})=1-\exp(-\|\mathbf{y}-\mathbf{x}\|_2^4)$. Thus, $\mathsf{L}_{t+}f(\mathbf{x})=0$ from (1.18), and hence from (1.22) it follows (1.19). Let $f(\mathbf{y})=(\mathbf{y}-\mathbf{x})\cdot\mathbf{z}$ on $K(\mathbf{x})$, which we extend to a bounded smooth function on $\mathbb{X}$. Thus, $\mathsf{L}_{t+}f(\mathbf{x})=\mathbf{b}(t,\mathbf{x})\cdot\mathbf{z}$ from (1.18), and hence from (1.22) it follows (1.20). Lastly, let $f(\mathbf{y})=\frac{1}{2}(\mathbf{y}-\mathbf{x})\otimes(\mathbf{y}-\mathbf{x})\cdot Z$ on $K(\mathbf{x})$, which we extend to a bounded smooth function on $\mathbb{X}$. Thus, $\mathsf{L}_{t+}f(\mathbf{x})=A(t,\mathbf{x})\cdot Z$ from (1.18), and hence from (1.22) it follows (1.21). ■

The condition (1.19) is about the path continuity (in probability): let $K(\mathbf{x})=\mathsf{B}_\epsilon(\mathbf{x})$, we have

$$\Pr(\|\mathsf{X}_{t+h}-\mathsf{X}_t\|\geq\epsilon|\mathsf{X}_t)=o(h),$$

i.e., X_{t+h} is unlikely to move too far away from X under small time increment h. Interestingly, the path continuity condition (1.19) is also equivalent to the "locality" of the generator L_{t+}, i.e., $\mathsf{L}_{t+}f(\mathbf{x})=0$ whenever f vanishes locally around $\mathbf{x}$. It is clear that (1.19) implies the continuity of the operator $\mathsf{T}_{s,t}$ in (1.15):

$$\forall f\in\mathcal{C}_b,\quad \lim_{t\downarrow s}\mathsf{T}_{s,t}f(\mathbf{x})=\mathsf{T}_{s,s}f(\mathbf{x})=f(\mathbf{x}),\ \textit{i.e.},\ \lim_{t\downarrow s}p(s,\mathbf{x},t,\mathrm{d}\mathbf{y})\to\delta_{\mathbf{x}}.$$

When $p(s,\mathbf{x},t,\mathrm{d}\mathbf{y})$ satisfies the following condition (e.g., Friedman 2006, Lemma 4.1, p. 114):

$$\exists\delta>0,\quad \lim_{h\downarrow 0}\tfrac{1}{h}\mathbb{E}[\|\mathsf{X}_{t+h}-\mathsf{X}_t\|^{2+\delta}|\mathsf{X}_t]=0,$$

then (1.19) automatically holds while (1.20)-(1.21) can be simplified to:

$$\lim_{h\downarrow 0}\frac{\mathbb{E}[\mathsf{X}_{t+h}|\mathsf{X}_t]-\mathsf{X}_t}{h}=\mathbf{b}(t,\mathsf{X}_t),\qquad \lim_{h\downarrow 0}\frac{\mathbb{E}[(\mathsf{X}_{t+h}-\mathsf{X}_t)\otimes(\mathsf{X}_{t+h}-\mathsf{X}_t)|\mathsf{X}_t]}{h}=A(t,\mathsf{X}_t).$$

A similar result holds for the (left) generator L_{s-}. Extension beyond the second order is apparent.

Skorokhod, A. V. (1996). "Lectures on the Theory of Stochastic Processes". De Gruyter.
Friedman, A. (2006). "Stochastic Differential Equations and Applications". Dover reprint.

Remark 1.69: Enforcing continuity

Suppose the convergence in the generators (1.16)-(1.17) is locally uniform in t and s, then L_{t+} and L_{s-} are right and left continuous, respectively, and moreover,

$$\mathsf{L}_{t-}=\lim_{s\uparrow t}\mathsf{L}_{s+},\qquad \mathsf{L}_{s+}=\lim_{t\downarrow s}\mathsf{L}_{t-}.$$

Additionally, if we know either L_{t+} or L_{t-} is continuous, then they coincide, which is signaled by the notation:

$$\mathsf{L}_t = \mathsf{L}_{t+} = \mathsf{L}_{t-}.$$

In particular, if the limits in (1.19)-(1.21) are locally uniform in t while b and A are continuous in t, then $\mathsf{L}_t = \mathsf{L}_{t+} = \mathsf{L}_{t-}$ exists.

For any $f \in \mathcal{C}_c^2$ (or even $\mathcal{C}_c^\infty$), if the limits in (1.19)-(1.21) are uniform in $\mathbf{x}$ while b and A are locally bounded in $\mathbf{x}$ (in particular, if they are continuous, so that $\mathsf{L}_{t+} f \in \mathcal{B}(\Sigma)$), then $\frac{\mathsf{T}_{t,t+h}-\mathrm{Id}}{h} f(\mathbf{x})$ converges uniformly in $\mathbf{x}$. Since $\mathsf{T}_{s,t} : \mathcal{B}(\Sigma) \to \mathcal{B}(\Sigma)$ is contractive (w.r.t. the sup norm), we have

$$\partial_{t+} \mathsf{T}_{s,t} f(\mathbf{x}) = \lim_{h \downarrow 0} \frac{\mathsf{T}_{s,t+h} - \mathsf{T}_{s,t}}{h} f(\mathbf{x}) = \lim_{h \downarrow 0} \mathsf{T}_{s,t} \frac{\mathsf{T}_{t,t+h} - \mathrm{Id}}{h} f(\mathbf{x}) = \mathsf{T}_{s,t} \lim_{h \downarrow 0} \frac{\mathsf{T}_{t,t+h} - \mathrm{Id}}{h} f(\mathbf{x}) = \mathsf{T}_{s,t} \mathsf{L}_{t+} f(\mathbf{x}),$$
$$\partial_{t+} \mathsf{T}_{s,t}^* = \mathsf{T}_{s,t}^* \mathsf{L}_{t+}^*.$$

Under similar conditions, we also have

$$\partial_{s-} \mathsf{T}_{s,t} = \mathsf{L}_{s-} \mathsf{T}_{s,t}, \qquad \partial_{s-} \mathsf{T}_{s,t}^* = \mathsf{L}_{s-}^* \mathsf{T}_{s,t}^*.$$

Combining the above observations, we know that when the limits in (1.19)-(1.21) are locally uniform in t and uniform in $\mathbf{x}$, and $\mathbf{b}$ and A are continuous in $(t, \mathbf{x})$, then for any $f \in \mathcal{C}_c^2$,

$$\boxed{\mathsf{L}_t f = \tfrac{1}{2} A \cdot \nabla^2 f + \mathbf{b} \cdot \nabla f, \quad \partial_t \mathsf{T}_{s,t} f = \mathsf{T}_{s,t} \mathsf{L}_t f, \quad \partial_s \mathsf{T}_{s,t} f = \mathsf{L}_s \mathsf{T}_{s,t} f.}$$

Moreover, using integration by parts, we have the adjoint relation:

$$\boxed{\mathsf{L}_t^* \mu = \tfrac{1}{2} \nabla^2 \cdot (\mu A) - \nabla \cdot (\mu \mathbf{b}), \quad \partial_t \mu \mathsf{T}_{s,t}^* = \mu \mathsf{T}_{s,t}^* \mathsf{L}_t^*, \quad \partial_s \mu \mathsf{T}_{s,t}^* = \mu \mathsf{L}_s^* \mathsf{T}_{s,t}^*.}$$

(Recall that $f \in \mathcal{C}_c^2$ so the boundary conditions vanish.) Alternatively, assume the limits in (1.19)-(1.21) are locally uniform in t, the function $u(s, \mathbf{x}) := \mathsf{T}_{s,t} \varphi(\mathbf{x}) \in \mathcal{C}_b^2$ (jointly in $(t, \mathbf{x})$ where boundedness follows from $\varphi \in \mathcal{B}(\Sigma)$), and $\mathbf{b}$ and A in (1.18) are continuous in t. Then, $\boxed{-\partial_s u(s, \mathbf{x}) = \mathsf{L}_s u(s, \mathbf{x})}$. This follows immediately from Theorem 1.68 by putting $f = u(s, \mathbf{x})$ (with s fixed) so that

$$\mathsf{L}_s u(s, \mathbf{x}) = \lim_{h \downarrow 0} \frac{\mathsf{T}_{s-h,s} - \mathrm{Id}}{h} u(s, \mathbf{x}) = \lim_{h \downarrow 0} \frac{u(s-h, \mathbf{x}) - u(s, \mathbf{x})}{h} = -\partial_s u(s, \mathbf{x}).$$

(Since the LHS is continuous in s, we know the left derivative on the RHS is indeed a derivative.)

Theorem 1.70: Fokker-Planck-Kolmogorov (FPK) equation (Kolmogorov 1931, 1933)

Let $p(s, \mathbf{x}, t, \mathbf{y})$ be (Markov) densities that satisfy (1.19)-(1.21) locally uniformly in t and uniformly in $\mathbf{x}$. Assume $\mathbf{b}$ and A in (1.18) are (jointly) continuous. Then,

$$\partial_t p(s, \mathbf{x}, t, \mathbf{y}) = \mathsf{L}_t^* p(s, \mathbf{x}, t, \mathbf{y}) = -\nabla_{\mathbf{y}} \cdot \big(p(s, \mathbf{x}, t, \mathbf{y}) \mathbf{b}(t, \mathbf{y})\big) + \tfrac{1}{2} \nabla_{\mathbf{y}}^2 \cdot \big(p(s, \mathbf{x}, t, \mathbf{y}) A(t, \mathbf{y})\big) \quad \textit{(forward)}$$
$$-\partial_s p(s, \mathbf{x}, t, \mathbf{y}) = \mathsf{L}_s p(s, \mathbf{x}, t, \mathbf{y}) = \mathbf{b}(s, \mathbf{x}) \cdot \nabla_{\mathbf{x}} p(s, \mathbf{x}, t, \mathbf{y}) + \tfrac{1}{2} A(s, \mathbf{x}) \cdot \nabla_{\mathbf{x}}^2 p(s, \mathbf{x}, t, \mathbf{y}) \quad \textit{(backward)},$$

provided that $\partial_t p(s, \mathbf{x}, t, \mathbf{y})$ is continuous in $(t, \mathbf{y})$, and for the backward equation, additionally $\nabla_{\mathbf{x}} p(s, \mathbf{x}, t, \mathbf{y})$ and $\nabla_{\mathbf{x}}^2 p(s, \mathbf{x}, t, \mathbf{y})$ are also continuous in $(\mathbf{x}, \mathbf{y})$.

Proof: Consider any $f \in \mathcal{C}_c^2$. Apply Theorem 1.68 we obtain

$$\langle f; \partial_t p(s, \mathbf{x}, t, \cdot) \rangle = \partial_t \langle f; p(s, \mathbf{x}, t, \cdot) \rangle = \mathsf{T}_{s,t} \mathsf{L}_t f(\mathbf{x}) = \langle \mathsf{L}_t f; p(s, \mathbf{x}, t, \cdot) \rangle = \langle f; \mathsf{L}_t^* p(s, \mathbf{x}, t, \cdot) \rangle$$
$$\langle f; -\partial_s p(s, \mathbf{x}, t, \cdot) \rangle = -\partial_s \langle f; p(s, \mathbf{x}, t, \cdot) \rangle = \mathsf{L}_s \mathsf{T}_{s,t} f(\mathbf{x}) = \mathsf{L}_s \langle f; p(s, \mathbf{x}, t, \cdot) \rangle = \langle f; \mathsf{L}_s p(s, \mathbf{x}, t, \cdot) \rangle.$$

(Note the minus sign for the backward equation, due to the definition of L_{s-} in Equation (1.17).) ■

Let us fix $s = 0$ and integrate $\mathbf{x}$ w.r.t. some initial distribution μ_0 (assuming we can push differentiation w.r.t. t under the integral w.r.t. $\mathbf{x}$):

$$\partial_t p(t, \mathbf{y}) = \mathsf{L}_t^* p(t, \mathbf{y}) = -\nabla \cdot (p\mathbf{b}) + \tfrac{1}{2}\nabla^2 \cdot (pA),$$

i.e., the marginal density also satisfies the forward equation.

See the monograph of Bogachev et al. (2015) for more on FPK.

Kolmogorov, A. N. (1931). "Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung". *Mathematische Annalen*, vol. 104. English translation at https://doi.org/10.1007/978-94-011-2260-3_9, pp. 415–458.
— (1933). "Zur Theorie der stetigen zufälligen Prozesse". *Mathematische Annalen*, vol. 108. English translation at https://doi.org/10.1007/978-94-011-2260-3_17, pp. 149–160.
Bogachev, V. I., N. V. Krylov, M. Röckner, and S. V. Shaposhnihov (2015). "Fokker-Planck-Kolmogorov Equations". American Mathematical Society.

Remark 1.71: Uniqueness

The uniqueness of a probability solution to the FPK equation has been settled by Bogachev et al. (2021a); see also Bogachev et al. (2020, 2021b) and Bogachev and Shaposhnihov (2021).

Bogachev, V. I., T. I. Krasovitskiy, and S. V. Shaposhnihov (2021a). "On uniqueness of probability solutions of the Fokker-Planck-Kolmogorov equation". *Sbornik: Mathematics*, vol. 212, no. 6, pp. 745–781.
— (2020). "The Kolmogorov Problem on Uniqueness of Probability Solutions of a Parabolic Equation". *Doklady Mathematics*, vol. 102, no. 3, pp. 464–467.
— (2021b). "On nonuniqueness of probability solutions to the Cauchy problem for the Fokker-Planck-Kolmogorov equation". *Doklady Mathematics*, vol. 103, no. 3, pp. 108–112.
Bogachev, V. I. and S. V. Shaposhnihov (2021). "Uniqueness of a Probability Solution to the Kolmogorov Equation with a Diffusion Matrix Satisfying Dini's Condition". *Doklady Mathematics*, vol. 104, no. 3, pp. 322–325.

History 1.72: Wolfgang Doeblin and his Pli Cacheté

See the documentary: Wolfgang Doeblin A Mathematician Rediscovered, and the excellent commentary (Bru and Yor 2002) as well as the original Pli (Doeblin 2000).

Bru, B. and M. Yor (2002). "Comments on the life and mathematical legacy of Wolfgang Doeblin Original Paper". *Finance and Stochastics*, vol. 6, pp. 3–47.
Doeblin, W. (2000). "Sur l'équation de Kolmogoroff". *Comptes Rendus de l'Académie des Sciences*, vol. 331, no. 12, pp. 1059–1102.

Theorem 1.73: FPK for SDE

Consider the stochastic differential equation

$$\mathrm{d}\mathsf{X}_t = \mathbf{b}_t(\mathsf{X}_t)\,\mathrm{d}t + G_t(\mathsf{X}_t)\,\mathrm{d}\mathsf{B}_t$$

and its generator

$$\mathsf{L}_t f(\mathbf{x}) := \mathbf{b}_t(\mathbf{x}) \cdot \nabla f(\mathbf{x}) + \tfrac{1}{2}[G_t(\mathbf{x})G_t(\mathbf{x})^\top] \cdot \nabla^2 f(\mathbf{x}).$$

Then, assuming $\partial_t p(s, \mathbf{x}, t, \mathbf{y})$ is continuous in $(t, \mathbf{y})$, we have

$$\partial_t p(s, \mathbf{x}, t, \mathbf{y}) = \mathsf{L}_t^* p(s, \mathbf{x}, t, \mathbf{y}) = -\nabla_{\mathbf{y}} \cdot [p(s, \mathbf{x}, t, \mathbf{y})\mathbf{b}_t(\mathbf{y})] + \tfrac{1}{2}\nabla_{\mathbf{y}}^2 \cdot \left[p(s, \mathbf{x}, t, \mathbf{y})G_t(\mathbf{y})G_t(\mathbf{y})^\top\right].$$

Proof: For any $f \in \mathcal{C}_c^2$, we apply Itô's formula to obtain

$$\mathrm{d}f(\mathsf{X}_t) = \left(\mathbf{b}_t(\mathsf{X}_t) \cdot \nabla f(\mathsf{X}_t) + \tfrac{1}{2}[G_t(\mathsf{X}_t)G_t^\top(\mathsf{X}_t)] \cdot \nabla^2 f(\mathsf{X}_t)\right)\mathrm{d}t + \nabla f(\mathsf{X}_t) \cdot G_t(\mathsf{X}_t)\,\mathrm{d}\mathsf{B}_t.$$

Take expectation from s at position $\mathbf{x}$ to t on both sides:

$$\mathrm{d}\mathbb{E}[f(\mathsf{X}_t)] = \mathbb{E}\left[\mathbf{b}_t(\mathsf{X}_t)\cdot\nabla f(\mathsf{X}_t) + \tfrac{1}{2}[G_t(\mathsf{X}_t)G_t^\top(\mathsf{X}_t)]\cdot\nabla^2 f(\mathsf{X}_t)\right]\mathrm{d}t$$
$$\partial_t\langle f; p(s,\mathbf{x},t,\mathbf{y})\rangle = \langle f; \partial_t p(s,\mathbf{x},t,\mathbf{y})\rangle = \langle \mathsf{L}_t f; p(s,\mathbf{x},t,\mathbf{y})\rangle = \langle f; \mathsf{L}_t^* p(s,\mathbf{x},t,\mathbf{y})\rangle\,,$$

where we have exchanged differentiation and integration in the first equality. ■

Example 1.74: Continuity equation

Consider the ordinary differential equation (i.e., setting $G_t \equiv \mathbf{0}$)

$$\mathrm{d}\mathsf{X}_t = \mathbf{b}_t(\mathsf{X}_t)\,\mathrm{d}t,$$

where the velocity field $\mathbf{b}_t$ models the movement per unit time. It follows from Theorem 1.73 that

$$\partial_t p(s,\mathbf{x},t,\mathbf{y}) = -\nabla_{\mathbf{y}}\cdot[p(s,\mathbf{x},t,\mathbf{y})\mathbf{b}_t(\mathbf{y})]\,.$$

Integrating w.r.t. some initial distribution at s (and exchanging differentiation with integration), we obtain the continuity equation (see also Liouville's Theorem):

$$\partial_t p_s(t,\mathbf{y}) = -\nabla_{\mathbf{y}}\cdot[p_s(t,\mathbf{y})\mathbf{b}_t(\mathbf{y})]\,,$$

where $p_s(t,\mathbf{y})$ is the marginal density at time t when we start from time s with some given initial distribution. Roughly, $\partial_t p_s(t,\mathbf{y})\,\mathrm{d}\mathbf{y}$ models the accumulation of quantity per unit volume ($\mathrm{d}\mathbf{y}$) and unit time, while $\nabla_{\mathbf{y}}\cdot[p_s(t,\mathbf{y})\mathbf{b}_t(\mathbf{y})]\,\mathrm{d}\mathbf{y}$, according to the divergence theorem, gives the difference between quantity flowing out and flowing in. Thus, the continuity equation states that the rate of accumulation of quantity exactly matches the difference between the in-flow rate and out-flow rate. [The product $p_s(t,\mathbf{y})\mathbf{b}_t(\mathbf{y})$ is called the flux, i.e., amount of quantity flowing per unit time, through a unit *area*.]

Example 1.75: Heat equation

Let us start with examining the heat equation

$$\partial_t p(t,\mathbf{x}) = \Delta_{\mathbf{x}} p(t,\mathbf{x}), \tag{1.23}$$

which corresponds to the (trivial) SDE

$$\mathrm{d}\mathsf{X}_t = \sqrt{2}\,\mathrm{d}\mathsf{B}_t.$$

(Note that $G_t = \sqrt{2}I$, not merely $\sqrt{2}$.) If we treat $p_t = p(t,\cdot) \in L_2(\mathrm{d}\mathbf{x})$, then we can rewrite the PDE (1.23) as an ODE with state space L_2:

$$\frac{\mathrm{d}p_t}{\mathrm{d}t} = -\nabla f(p_t),$$

where the energy function $f(p) := \frac{1}{2}\|\nabla p\|_{L_2}^2$. Indeed,

$$\frac{\mathrm{d}f(p+\epsilon q)}{\mathrm{d}\epsilon}\Big|_{\epsilon=0} = \int \nabla p\cdot\nabla q\,\mathrm{d}\mathbf{x} = -\int q\Delta p\,\mathrm{d}\mathbf{x}.$$

Example 1.76: Continuous flow for density estimation (Tabak and Vanden-Eijnden 2010)

Let p be a known density, e.g., standard normal, and q be a density we wish to learn. We know there exist mappings T so that

$$q = \mathsf{T}_{\#}^{-1} p, \quad i.e., \quad \mathsf{T}^{-1}(\mathsf{X}) \sim q \text{ if } \mathsf{X} \sim p.$$

Let us build T^{-1} through a continuous family T_t^{-1}, leading to

$$p_t := (\mathsf{T}_t^{-1})_{\#} p, \quad p_t(\mathbf{x}) = |\det \mathsf{T}_t' \mathbf{x}| \cdot p(\mathsf{T}_t \mathbf{x}).$$

Conversely, we also have

$$p = \mathsf{T}_{\#} q, \quad q_t := (\mathsf{T}_t)_{\#} q, \quad q_t(\mathsf{T}_t \mathbf{x}) \cdot |\det \mathsf{T}_t' \mathbf{x}| = q(\mathbf{x}).$$

We use the KL divergence as our objective of learning:

$$\mathsf{KL}(q \| p_t) = \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p_t(\mathbf{x})} \, \mathrm{d}\mathbf{x} = \mathsf{KL}(q_t \| p) = \int q(\mathbf{x}) \log \frac{q_t(\mathsf{T}_t \mathbf{x})}{p(\mathsf{T}_t \mathbf{x})} \, \mathrm{d}\mathbf{x}.$$

We take the (functional) derivative w.r.t. T_t:

$$\frac{\delta \mathsf{KL}}{\delta \mathsf{T}_t}(\mathbf{x}) = [\mathbf{s}_{q_t}(\mathsf{T}_t \mathbf{x}) - \mathbf{s}_p(\mathsf{T}_t \mathbf{x})] q(\mathbf{x}) = [\mathbf{s}_{q_t}(\mathsf{T}_t \mathbf{x}) - \mathbf{s}_p(\mathsf{T}_t \mathbf{x})] \cdot q_t(\mathsf{T}_t \mathbf{x}) \cdot |\det \mathsf{T}_t' \mathbf{x}| \tag{1.24}$$

and evolve T_t according to the ODE (that guarantees decrease of our KL objective):

$$\mathrm{d}\mathsf{T}_t = -\mathbf{b}(\mathsf{T}_t), \quad \text{where} \quad \mathbf{b}(\mathbf{z}) = [\mathbf{s}_{q_t}(\mathbf{z}) - \mathbf{s}_p(\mathbf{z})] \cdot q_t(\mathbf{z}). \tag{1.25}$$

We have dropped the Jacobian $|\det \mathsf{T}_t' \mathbf{x}|$ in (1.24) for better interpretation. Essentially, we seek an infinitesimal improvement T over our current estimate T_t, so we compute $\delta \mathsf{KL}(\mathsf{T} \circ \mathsf{T}_t) / \delta \mathsf{T} \restriction_{\mathsf{T}=\mathrm{Id}}$, which leads exactly to (1.25). This Lagrangian view allows us to "forget" the past and focus "myopically" on deforming the *current* q_t to the target p. In fact, we have the following continuity equation corresponding to (1.25):

$$\partial_t q_t = \nabla \cdot (q_t \mathbf{b}) = \nabla \cdot [q_t^2 (\mathbf{s}_{q_t} - \mathbf{s}_p)].$$

Alternatively, we may use the reverse KL divergence as our objective:

$$\mathsf{LK}(q \| p_t) = \mathsf{LK}(q_t \| p) = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q_t(\mathbf{x})} \, \mathrm{d}\mathbf{x}.$$

Interestingly, the dynamics (1.25) also decreases the reverse KL divergence:

$$\begin{aligned} \frac{\mathrm{d}\mathsf{LK}}{\mathrm{d}t} &= -\int \frac{p(\mathbf{x})}{q_t(\mathbf{x})} \partial_t q_t(\mathbf{x}) \, \mathrm{d}\mathbf{x} = -\int \frac{p(\mathbf{x})}{q_t(\mathbf{x})} \nabla \cdot [q_t^2 (\mathbf{s}_{q_t} - \mathbf{s}_p)] \, \mathrm{d}\mathbf{x} = \int [q_t^2 (\mathbf{s}_{q_t} - \mathbf{s}_p)] \cdot \nabla \frac{p(\mathbf{x})}{q_t(\mathbf{x})} \, \mathrm{d}\mathbf{x} \\ &= -\int p q_t \|\mathbf{s}_{q_t} - \mathbf{s}_p\|_2^2 \, \mathrm{d}\mathbf{x}. \end{aligned}$$

Tabak, E. G. and E. Vanden-Eijnden (2010). "Density estimation by dual ascent of the log-likelihood". *Communications in Mathematical Sciences*, vol. 8, no. 1, pp. 217–233.

Example 1.77: Score matching as gradient flow of KL (Lyu 2009)

The Fisher divergence between two densities is the square L_2 distance between their score functions:

$$\mathsf{F}(p \| q) := \tfrac{1}{2} \|\mathbf{s}_p - \mathbf{s}_q\|_2^2.$$

Define the noisy versions

$$p_t = p * \mathcal{N}(0, t), \quad q_t = q * \mathcal{N}(0, t).$$

Lyu (2009) proved that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}(p_t\|q_t) = -\mathsf{F}(p_t\|q_t).$$

Lyu, S. (2009). "Interpretation and generalization of score matching". In: *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 359–366.

Example 1.78: Neural ODE (Chen et al. 2018)

Chen, T. Q., Y. Rubanova, J. Bettencourt, and D. K. Duvenaud (2018). "Neural ordinary differential equations". In: *Advances in Neural Information Processing Systems*, pp. 6572–6583.

Example 1.79: Variational optimal transport (Benamou and Brenier 2000)

Let p_0 and p_1 be two given densities, with Wasserstein distance

$$\mathbb{W}_2^2(p_0,p_1) = \min_{\mathsf{T}:\mathsf{T}_{\#}p_0=p_1} \mathbb{E}_{\mathsf{X}\sim p_0}\|\mathsf{T}\mathsf{X}-\mathsf{X}\|_2^2, \quad \text{where} \quad p_0(\mathbf{x}) = p_1(\mathsf{T}\mathbf{x})\cdot\det\mathsf{T}'\mathbf{x}.$$

Brenier (1991) proved the existence of such a (unique) $\mathsf{T}=\nabla\varphi$ for some convex potential function φ.

Let us consider the ODE:

$$\mathrm{d}\mathsf{X}_t = \mathbf{b}_t(\mathsf{X}_t),$$

where $\mathsf{X}_0\sim p_0$ and $\mathsf{X}_1\sim p_1$. Obviously, this gives us a way to interpolate densities: $\mathsf{X}_t\sim p_t$, which satisfies the continuity equation:

$$\partial_t p_t = -\nabla\cdot(p_t\mathbf{b}_t).$$

In fact, Benamou and Brenier (2000) showed the following surprising result:

$$\int\int_0^1 p_t(\mathbf{x})\|\mathbf{b}_t(\mathbf{x})\|_2^2\,\mathrm{d}\mathbf{x}\,\mathrm{d}t = \int\int_0^1 p_0(\mathbf{x})\|\mathbf{b}_t(\mathsf{X}_t)\|_2^2\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$$

See also Benamou and Brenier 2001; Guittet 2003.

Benamou, J.-D. and Y. Brenier (2000). "A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem". *Numerische Mathematik*, vol. 84, pp. 375–393.
Brenier, Y. (1991). "Polar factorization and monotone rearrangement of vector-valued functions". *Communications on Pure and Applied Mathematics*, vol. 44, no. 4, pp. 375–417.
Benamou, J.-D. and Y. Brenier (2001). "Mixed L_2-Wasserstein Optimal Mapping Between Prescribed Density Functions". *Journal of Optimization Theory and Applications*, vol. 111, pp. 255–271.
Guittet, K. (2003). "On the Time-Continuous Mass Transport Problem and Its Approximation by Augmented Lagrangian Techniques". *SIAM Journal on Numerical Analysis*, vol. 41, no. 1, pp. 382–399.

Example 1.80: FPK as gradient flow (Jordan et al. 1998)

Consider the SDE

$$\mathrm{d}\mathsf{X}_t = -\nabla f(\mathsf{X}_t)\,\mathrm{d}t + \sqrt{2\beta^{-1}}\,\mathrm{d}\mathsf{B}_t,$$

and the FPK equation

$$\partial_t p = \nabla\cdot(p\nabla f) + \beta^{-1}\Delta p = \nabla\cdot[p(\nabla f + \beta^{-1}s)],$$

where $s(t, \mathbf{x}) = \nabla_{\mathbf{x}} \log p(t, \mathbf{x})$ is the score function. When the potential f satisfies appropriate growth conditions, there is a unique stationary solution of FPK (the so-called Gibbs distribution):

$$s = -\beta \nabla f \iff p \propto \exp(-\beta f).$$

Jordan, R., D. Kinderlehrer, and F. Otto (1998). "The Variational Formulation of the Fokker–Planck Equation". *SIAM Journal on Mathematical Analysis*, vol. 29, no. 1, pp. 1–17.