

Gradient Flows and Marginal Maximum Likelihood Estimation

From EM and ULA to PGD and back again

Adam M. Johansen (University of Warwick)

Joint work with Rocco Caprio, Jen Ning Lim, Juan Kuntz and Sam Power

OCEAN Retreat—September 1st, 2026

a.m.johansen@warwick.ac.uk

<http://go.warwick.ac.uk/amjohansen/talks/>



Gradient flows and related analytic techniques are useful:

- Developing novel algorithms
- Understanding existing ones

Outline:

- The EM Algorithm for MMLE
- Gradient flows and hybrid spaces
 - Idea
 - The PGD Algorithm
 - A little non-asymptotic theory
- Something this tells us about EM

Euclidean Given $f : \mathbb{R}^d \rightarrow \mathbb{R}$, the ODE

$$dx_t = -\nabla f(x_t)dt, x_0 \in \mathbb{R}^d$$

admits time discretisation

$$x_{t+h} = x_t + h\nabla f(x_t).$$

General Given $F : \mathcal{X} \rightarrow \mathbb{R}$, we envisage

$$dX_t = -\text{grad}_{\mathcal{X}} dt, X_0 \in \mathcal{X}$$

and some time discretisation thereof.

Requires an interpretation of $\text{grad}_{\mathcal{X}}$; easy on Riemannian manifolds and more generally via “Otto calculus” (Otto, 2001; Villani, 2009).

Inference in Latent Variable Models

Problem: Given some (fixed) data y , and probabilistic model $p_\theta(x, y)$ relating latent variables x and data y , parametrized by θ : infer θ and characterise $p_\theta(x|y)$.

Assumption: x and θ take values in $\mathbb{R}^{d_x}$ and $\mathbb{R}^{d_\theta}$, respectively.

A Simple Example

Often, data is assumed to be generated as

$$y_m = f_\theta(x_m) + \eta_m, \quad x_m \sim p_\theta^x(dx), \quad \forall m = 1, \dots, M;$$

where $\{\eta_m\}$ are i.i.d. Gaussian noises. In this case,

$$p_\theta(x, y) = \prod_{m=1}^M p^\eta(y_m - f_\theta(x_m)) p_\theta^x(x_m).$$

Marginal Maximum Likelihood / Empirical Bayes

Problem: Given some (fixed) data y , and probabilistic model $p_\theta(x, y)$ relating latent variables x and data y , parametrized by θ : infer θ and characterise $p_\theta(x|y)$.

Approach: Empirical Bayes (Robbins, 1956): search for a θ_* that explains the data well, and use that to infer x :

1. Find θ_* maximizing the marginal likelihood,

$$\int p_\theta(x, y) dx =: p_\theta(y):$$

$$\theta_* := \arg \max_{\theta} \int p_\theta(x, y) dx.$$

2. Use θ_* to obtain

$$p_{\theta_*}(x|y) = \frac{p_{\theta_*}(x, y)}{p_{\theta_*}(y)}$$

and perform inference on x via this ‘optimal posterior’.

The EM algorithm

The canonical approach to this problem is the **EM algorithm** (Dempster et al., 1977):

Given some (θ_0, q_0) , iterate

- M-step

$$\theta_{k+1} = \arg \max_{\theta} \int \log(p_{\theta}(x, y)) q_k(dx)$$

- E-step

$$q_{k+1} = p_{\theta_{k+1}}(\cdot|y)$$

The EM algorithm as optimization

Proposition (Neal and Hinton (1998))

EM is a **coordinate-wise minimization** procedure on

$\mathcal{M} := \mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})$:

$$\theta_{k+1} \in \operatorname{argmin}_{\theta \in \mathbb{R}^{d_\theta}} \mathcal{F}(\theta, q_k),$$

$$q_{k+1} \in \operatorname{argmin}_{q \in \mathcal{P}(\mathbb{R}^{d_x})} \mathcal{F}(\theta_{k+1}, q),$$

where $\mathcal{F} : \mathbf{R}^{d_\theta} \times \mathcal{P}(\mathbf{R}^{d_x}) \rightarrow \mathbf{R}$ is the *free energy* functional:

$$\mathcal{F}(\theta, q) := \int \log \left(\frac{q(x)}{p_\theta(x, y)} \right) q(dx)$$

and

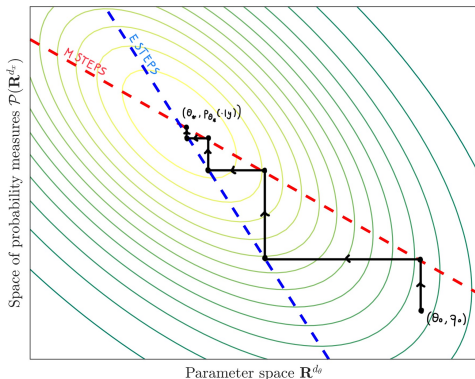
$$\mathcal{M}_\star := \operatorname{argmin} F = \{(\theta_\star, p_{\theta_\star}(\cdot|y)) : \theta_\star \in \arg \max p_\theta(y)\}.$$

The EM algorithm as optimization

$$\theta_{k+1} = \arg \max_{\theta} \int \log(p_{\theta_k}(x, y)) q_k(dx) = \operatorname{argmin}_{\theta \in \mathbb{R}^{d_{\theta}}} \mathcal{F}(\theta, q_k) \quad (\text{M})$$

$$q_{k+1} = p_{\theta_{k+1}}(\cdot | y) = \operatorname{argmin}_{q \in \mathcal{P}(\mathbb{R}^{d_x})} \mathcal{F}(\theta_{k+1}, q) \quad (\text{E})$$

Hence, EM does something like



The EM algorithm as optimization

What about other optimization approaches for $\mathcal{F}(\theta, q)$?

	$\theta \in \mathbb{R}^{d_\theta}$	$q \in \mathcal{P}(\mathbb{R}^{d_x})$
EM	min	min
Gradient EM	gradient step	min
Langevin EM	min	gradient step
Coordinate-wise Gradient Descent	gradient step	gradient step
Gradient Descent	gradient step	

To make sense of this table, we want to understand what it means to take gradient steps of $\mathcal{F}$ with respect to $(\theta, q) \in \mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})$.

Geometry of $\mathcal{P}(\mathbb{R}^{d_x})$

When endowed with the Wasserstein-2 metric d_{W_2} , $\mathcal{P}(\mathbb{R}^{d_x})$ can be thought of as a Riemannian manifold (Otto, 2001; Villani, 2009). Then, for some reasonable functions $\mathcal{G}$ on $\mathcal{P}(\mathbb{R}^{d_x})$ one can make sense of things like

- Gradient of $\mathcal{G}$ in $\mathcal{P}(\mathbb{R}^{d_x})$: $\text{grad}\mathcal{G}(q) = \nabla_x \cdot (q \nabla_x \delta_{\mathcal{P}_2} \mathcal{G}(q))$
- The gradient flow of probability distributions

$$\partial_t q_t = -\text{grad}\mathcal{G}(q_t): \partial_t q_t = \nabla_x \cdot (q_t \nabla_x \delta_{\mathcal{P}_2} \mathcal{G}(q_t)), \quad q_0 = q.$$

where $\delta_{\mathcal{P}_2} \mathcal{G}(q)$ is the functional derivative (first variation) of $\mathcal{G}$.

- If $\mathcal{G}(q) = \mathcal{H}(q) = \int q \log q$, then $\partial_t q_t = \Delta_x q_t$, the heat flow.
- If $\mathcal{G}(q) = \text{KL}(q||\pi) = \int q \log(q/\pi)$, $\pi \propto e^{-V}$, then

$\partial_t q_t = \nabla_x \cdot (\nabla_x q_t + q_t \nabla_x V)$, the Langevin diffusion's law:

$$dX_t = -\nabla_x V(X_t)dt + \sqrt{2}dW_t \quad X_t \sim q_t.$$

Geometry of $\mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})$

When endowed with the product-metric

$$d((\theta, q), (\theta', q')) = \sqrt{d_E(\theta, \theta')^2 + d_{W_2}(q, q')^2}$$

$(\mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x}))$ can be thought of as a Riemannian manifold
(Kuntz et al., 2023; Caprio et al., 2025)

This allow us to make sense of concepts such as

- Gradient of the free energy $\mathcal{F}$ in $\mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})$:

$$\text{grad}\mathcal{F}(\theta, q) = \left(- \int \nabla_\theta \log(p_\theta(x, y)) q(dx), -\nabla_x \cdot \{q(\nabla_x \log(q) - \nabla_x \log(p_\theta(\cdot, y)))\} \right).$$

- The gradient flow of the free energy $\partial_t(\theta_t, q_t) = -\text{grad}\mathcal{F}(\theta_t, q_t)$:

$$\partial_t \theta_t = \int \nabla_\theta \log(p_{\theta_t}(x, y)) q_t(dx), \quad \partial_t q_t = \nabla_x \cdot \{q_t(\nabla_x \log(q_t) - \nabla_x \log((p_{\theta_t}(\cdot, y)))\}$$

i.e.,

$$\partial_t \theta_t = \int \nabla_\theta \log(p_{\theta_t}(x, y)) \text{Law}(X_t)(dx), \quad dX_t = \nabla_x \log(p_{\theta_t}(X_t, y)) dt + \sqrt{2} dW_t$$

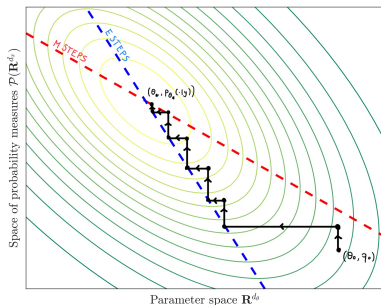
i.e. the *steepest descent curve* connecting an arbitrary (θ_0, q_0) to $(\theta_*, p_{\theta_*}(\cdot|y))$. 11

The EM algorithm as optimization: gradient EM

	$\theta \in \mathbb{R}^{d_\theta}$	$q \in \mathcal{P}(\mathbb{R}^{d_x})$
Gradient EM	gradient step	min

$$\theta_{k+1} = \theta_k + h \int \nabla_{\theta} \log(p_{\theta_k}(x, y)) q_k(dx) \quad (\tilde{M})$$

$$q_{k+1} = p_{\theta_{k+1}}(\cdot | y) \quad (E)$$

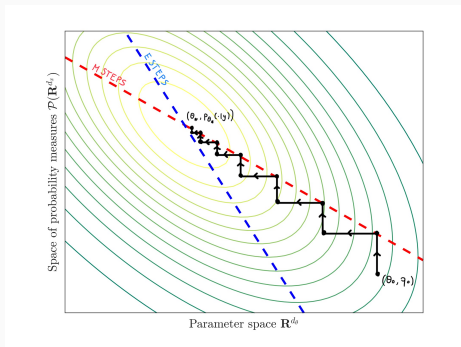


The EM algorithm as optimization: Langevin EM

	$\theta \in \mathbb{R}^{d_\theta}$	$q \in \mathcal{P}(\mathbb{R}^{d_x})$
Langevin EM	min	gradient step

$$\theta_{k+1} = \arg \max_{\theta} \int \log(p_{\theta_k}(x, y)) q_k(dx) \quad (\text{M})$$

$$q_{k+1} = \text{Law}(X_k + h \nabla_x \log(p_{\theta_{k+1}}(X_k|y)) + \sqrt{2h} \mathcal{N}_k), \quad \text{Law}(X_k) = q_k \quad (\tilde{\text{E}})$$

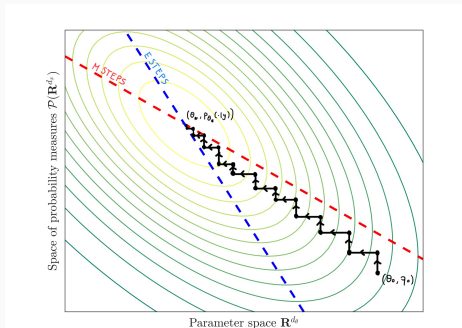


The EM algorithm as optimization: CW Gradient Descent

	$\theta \in \mathbb{R}^{d_\theta}$	$q \in \mathcal{P}(\mathbb{R}^{d_x})$
Coordinate-wise Gradient Descent	gradient step	gradient step

$$\theta_{k+1} = \theta_k + h \int \nabla_{\theta} \log(p_{\theta_k}(x, y)) q_k(dx) \quad (\tilde{M})$$

$$q_{k+1} = \text{Law}(X_k + h \nabla_x \log(p_{\theta_{k+1}}(X_k|y)) + \sqrt{2h} \mathcal{N}_k), \quad \text{Law}(X_k) = q_k \quad (\tilde{E})$$



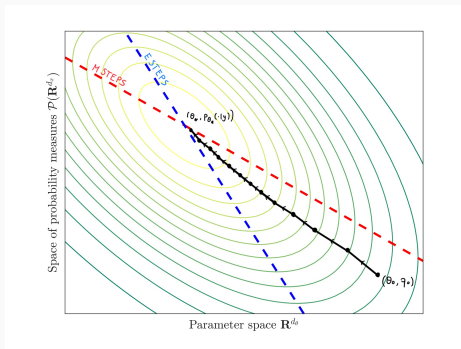
The EM algorithm as optimization: Gradient Descent

	$\theta \in \mathbb{R}^{d_\theta}$ $q \in \mathcal{P}(\mathbb{R}^{d_x})$
Gradient Descent	gradient step

$$\theta_{k+1} = \theta_k + h \int \nabla_{\theta} \log(p_{\theta_k}(x, y)) q_k(dx)$$

$$q_{k+1} = \text{Law}(X_k + h \nabla_x \log(p_{\theta_k}(X_k|y)) + \sqrt{2h} \mathcal{N}_k), \quad \text{Law}(X_k) = q_k$$

($\tilde{M}$ - $\tilde{E}$)



Particle gradient descent (Kuntz et al., 2023)

Discretizing the gradient flow, we obtain **particle gradient descent**:

(1) Choose the step size $h > 0$ and particle number $N > 0$:

$$\Theta_{k+1} = \Theta_k + \frac{h}{N} \sum_{n=1}^N \nabla_{\theta} \ell(\Theta_k, X_k^n), \quad \forall k \in [K] := \{1, \dots, K\},$$

$$X_{k+1}^n = X_k^n + h \nabla_x \ell(\theta_k, X_k^n) + \sqrt{2h} W_k^n, \quad \forall n \in [N], \quad k \in [K],$$

where $(W_k^n)_{k \in [K-1], n \in [N]}$ denote iid standard normal r.v.s., so

$$\Theta_k \approx \theta_{kh}, \quad \frac{1}{N} \sum_{n=1}^N \delta_{X_k^n} \approx q_{kh}, \quad \forall k > 0.$$

(2) Estimate a stationary point θ_* of the marginal likelihood $\theta \mapsto p_{\theta}(y)$ and its corresponding posterior $p_{\theta_*}(\cdot|y)$ using

$$\theta_* \approx \Theta_K, \quad p_{\theta_*}(\cdot|y) \approx \frac{1}{N} \sum_{n=1}^N \delta_{X_K^n}.$$

Error bounds via functional inequalities

1. We want to study non-asymptotics for the above algorithms.
2. They are, in a way or another, different methods to minimise $\mathcal{F}$, and their convergence will depend on properties of $\mathcal{F}$.
3. Furthermore, this suggests we should first study their convergence in terms of free energy $\mathcal{F}$.
4. For instance, for EM

- M-step

$$\theta_{k+1} = \arg \max_{\theta} \int \log(p_{\theta_k}(x, y)) q_k(dx)$$

- E-step

$$q_{k+1} = p_{\theta_{k+1}}(\cdot | y)$$

this suggests we should study the *joint* convergence of both θ_k and q_k in terms of free energy, $\mathcal{F}(\theta_k, q_k)$.

Error bounds via functional inequalities (Caprio et al., 2025)

We introduce a functional inequality that relates $\mathcal{F}$ with its *gradient*. Let

$$\mathcal{I}(\theta, q) := \left\| \int \nabla_{\theta} \log(p_{\theta}(x, y)) q(x) dx \right\|^2 + \int \left\| \nabla_x \log \left(\frac{q(x)}{p_{\theta}(x, y)} \right) \right\|^2 q(x) dx.$$

The functional $\mathcal{I}$ can be interpreted as $\|\text{grad}\mathcal{F}\|^2$, where $\text{grad}\mathcal{F}$ is the gradient of the free energy in the geometry induced by the Euclidean and Wasserstein metrics on $\mathbb{R}^{d_{\theta}} \times \mathcal{P}_2(\mathbb{R}^{d_x})$.

Definition (Extended log-Sobolev inequality; xLSI)

The statistical model $(p_{\theta}(x, y))_{\theta \in \mathbb{R}^{d_{\theta}}}$ satisfies the extended log-Sobolev inequality with constant $\lambda > 0$ if

$$2\lambda[\mathcal{F}(\theta, q) - \inf \mathcal{F}] \leq \mathcal{I}(\theta, q) \quad (\text{xLSI})$$

holds for all $(\theta, q) \in \mathcal{M}_2$.

Error bounds via functional inequalities

A few remarks on the xLSI:

1. If, in a sense, $\mathbb{R}^{d_\theta} = \{\theta\}$, this is the log-Sobolev inequality for $p_\theta(\cdot|y)$:

$$2\lambda \text{KL}(q||p_\theta(\cdot|y)) \leq \int \left\| \nabla_x \log \left(\frac{q(x)}{p_\theta(x|y)} \right) \right\|^2 q(x) dx;$$

2. If, in a sense, $\mathcal{P}_2(\mathbb{R}^{d_x}) = \{p_\theta(\cdot|y)\}$, this is the Polyak–Łojasiewicz inequality for $f(\theta) := -\log(p_\theta(y))$:

$$2\lambda[f(\theta) - \inf f] \leq \|\text{grad} f(\theta)\|^2;$$

3. The xLSI is a property of a *statistical model* / parametrized family of distributions.

Error bounds via functional inequalities

Just like the extended LSI generalises the standard log-Sobolev inequality, we have generalisations of various other key results and inequalities:

- An ‘extended’ Talagrand (1996) inequality:

$$\lambda d((\theta, q), \mathcal{M}_\star)^2 \leq 2(\mathcal{F}(\theta, q) - \mathcal{F}_\star), \quad \mathcal{M}_\star := \operatorname{argmin} \mathcal{F} \quad (1)$$

- An ‘extended’ Otto and Villani (2000) Theorem:

xLSI $\Rightarrow$ the extended Talagrand inequality.

Error bounds via functional inequalities: nice to have

- An ‘extended’ HWI inequality (Otto and Villani, 2000): if

$$\nabla^2 \log(p_\theta(x, y)) \preceq -\iota I_{d_\theta + d_x} \text{ for some } \iota \in \mathbb{R},$$

$$\mathcal{F}(\theta, q) \leq \mathcal{F}(\theta', q') + d((\theta, q), (\theta', q')) \sqrt{\mathcal{I}(\theta, q)} - \frac{\iota}{2} d((\theta, q), (\theta', q'))$$

- An extended Bakry and Émery (1985) Theorem: if

$$\nabla^2 \log(p_\theta(x, y)) \preceq -\iota I_{d_\theta + d_x} \text{ with } \iota > 0, \text{ the xLSI holds with } \lambda = \iota > 0.$$

- Preservation results: if the model $(p_\theta(x, y))_{\theta \in \Theta}$ satisfy the xLSI with constant λ , and “ $\tilde{p}_\theta(x, y)$ is some perturbation of $p_\theta(x, y)$ ” then also $(\tilde{p}_\theta(x, y))_{\theta \in \Theta}$ satisfy the extended log-Sobolev inequality with some explicit constant $\tilde{\lambda}$ (generalisation of things like Holley and Stroock (1987)’s lemma ...)

Error bounds via functional inequalities

Two assumptions we will consider (separately, or together).

We say the model $(p_\theta(x, y))_{\theta \in \mathbb{R}^{d_\theta}}$ is

1. L_θ -smooth if

$$\|\nabla_\theta \log(p_\theta(x, y)) - \nabla_\theta \log(p_{\theta'}(x, y))\| \leq L_\theta \|\theta - \theta'\|;$$

2. L_x -smooth if

$$\|\nabla_x \log(p_\theta(x, y)) - \nabla_x \log(p_\theta(x', y))\| \leq L_x \|x - x'\|;$$

and set $L := \max(L_\theta, L_x)$.

Theorem ((Caprio and Johansen, 2025))

If statistical model $(p_{\theta}(x, y))_{\theta \in \mathbb{R}^{d_{\theta}}}$ satisfy the xLSI + smoothness:

- For gradient descent:

$$\lambda d((\theta_k, q_k), \mathcal{M}_{\star})^2 \leq 2e^{-\lambda \sum_{j=1}^k h_j} (\mathcal{F}(\theta_0, q_0) - \mathcal{F}_{\star}) + 12L^2 d_x \sum_{j=1}^k h_j^2 e^{-\lambda \sum_{i=j+1}^k h_i}.$$

and PGD can also be handled (Caprio et al., 2025).

- For alternating minimisation (EM):

$$\lambda d((\theta_k, q_k), \mathcal{M}_{\star})^2 \leq 2 \inf \left\{ e^{-\lambda(k/L_{\theta} + \sum_{j=1}^k h_j)} (\mathcal{F}(\theta_0, q_0) - \mathcal{F}_{\star}) + B \sum_{j=1}^k h_j^2 e^{-\lambda(k/L_{\theta} + \sum_{i=j+1}^k h_i)} \right\}$$

with all $h_j \leq L_x/4$.

Conclusions

- One can take inspiration from optimisation (and sampling) to solve the maximum marginal likelihood estimation problem, and to study related algorithms theoretically;
- Connections between EM (and its friends) with functional inequalities run deep.
- Functional analysis; gradient flows and related topics still have more to tell us about computational statistics.

References

- Ö. Deniz Akyildiz, Francesca R. Crucinio, Mark Girolami, Tim Johnston, and Sotirios Sabanis. Interacting particle Langevin algorithm for maximum marginal likelihood estimation. *ESAIM: Probability and Statistics (to appear)*, 2025.
- Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84*. Lecture Notes in Mathematics Springer, 1985.
- Rocco Caprio and Adam M Johansen. Fast convergence of the expectation-maximization algorithm under a logarithmic sobolev inequality. *Biometrika*, 112(4):asaf061, 2025. doi: 10.1093/biomet/asaf061.
- Rocco Caprio, Juan Kuntz, Samuel Power, and Adam M Johansen. Error bounds for particle gradient descent, and extensions of the log-Sobolev and Talagrand inequalities. *Journal of Machine Learning Research*, 26(103):1–38, 2025. URL <https://jmlr.org/papers/v26/24-0437.html>.
- Valentin De Bortoli, Alain Durmus, Marcelo Pereyra, and Ana F Vidal. Efficient stochastic optimisation by unadjusted Langevin Monte Carlo: Application to maximum marginal likelihood and empirical Bayesian estimation. *Statistics and Computing*, 31: 1–18, 2021.
- Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, 39:2–38, 1977.
- Richard Holley and Daniel W Stroock. Logarithmic sobolev inequalities and stochastic ising models. *Journal of Statistical Physics*, 46:1159–1194, 1987.
- Juan Kuntz, Jen Ning Lim, and Adam M. Johansen. Particle algorithms for maximum likelihood training of latent variable models. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 5134–5180, 2023.
- Radford M. Neal and Geoffrey E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pages 355–368. Springer Netherlands, 1998.
- Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 26: 101–174, 2001.
- Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173:361–400, 2000.
- Herbert Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 3.1, pages 157–164, 1956.
- Michel Talagrand. Transportation cost for Gaussian and other product measures. *Geometric & Functional Analysis*, 6: 587–600, 1996.
- Cédric Villani. *Optimal Transport: Old and New*. Springer Science & Business Media, 2009.