Computing Wasserstein barycenters using gradient descent algorithms.

joint with Sinho Chewi, Tyler Maunu & Austin Stromme (MIT)

1. Averages

$X_1, \dots, X_n \overset{iid}{\sim} P$ over $\mathbb{R}$,

$$\hat{\theta}_n = \overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \qquad \theta = \mathbb{E}[X_1]$$

- LLN: $\hat{\theta}_n \to \theta$
- Quadratic risk: $\mathbb{E}[(\overline{X}_n - \mathbb{E}[X_1])^2] = \frac{\mathrm{var}(P)}{n}$

is a standard statistical technique

→ Regression, Maximum likelihood, empirical risk minimization, ...

Note that $\operatorname{argmin}_m \frac{1}{n}\sum_{i=1}^n (X_i - m)^2$ & $\operatorname{argmin}_m \mathbb{E}[(X_1 - m)^2]$ } both metric quantities

$$\overline{X}_n = \operatorname*{argmin}_m \frac{1}{n}\sum_{i=1}^n d^2(X_i, m)$$

$$\mathbb{E}[X_1] = \operatorname*{argmin}_m \int d^2(x, m)\, dP(x)$$

↳

2. Wasserstein barycenters

This talk: $(\mathbb{R}, |\cdot|) \rightsquigarrow (\mathcal{P}_2(\mathbb{R}^d), W_2) =: \mathcal{W}_2$

Wasserstein space.

> Captures interesting geometric features in graphics and Machine learning

Setup:
- P probability measure over $\mathcal{W}_2$
- $b^* = \operatorname*{argmin}_{b \in \mathcal{W}_2} \int W_2^2(b, \mu)\, dP(\mu)$

  ↳ true barycenter
- Observe $\mu_1, \dots, \mu_n \overset{iid}{\sim} P$ (images, datasets, ...)

$$\hat{b} = \operatorname*{argmin}_{b \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} W_2^2(b, \mu_i)$$

↳ empirical barycenter

Wasserstein barycenters (Agueh & Carlier 2011)

Questions: ✓ Statistical: $\mathbb{E}\, W_2^2(\hat{b}, b^*) \lesssim \frac{\sigma^2}{n}$?

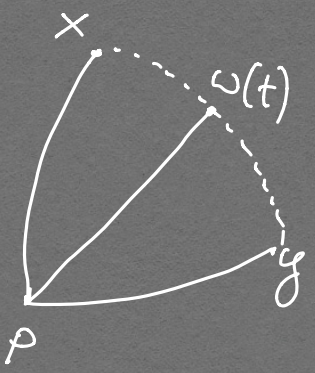
✓ Computational: - How to compute $\hat{b}$

- " " " "

at $\mathbb{E}\, W_2^2(\tilde{b}, b^*) \lesssim \frac{\sigma^2}{n}$

Why different from Euclidean case?

Main answer: CURVATURE



Def: $\mathcal{M} = (\Pi, d)$ geodesic space
$\mathrm{curv}(\mathcal{M}) \geq 0$ if ....

if $d(x,y) = \|x - y\|$ ← Hilbert norm

$$d^2(\rho, \omega(t)) = (1-t)\, d^2(\rho, x) + t\, d^2(\rho, y) - t(1-t)\, d^2(x,y)$$

$\forall\ \rho, x, y$

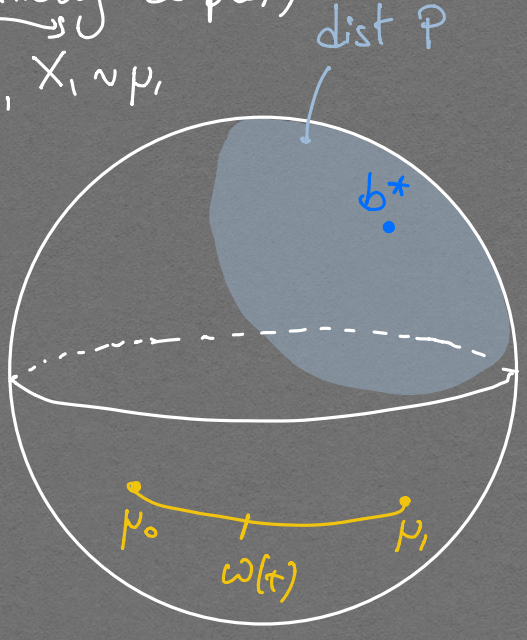
geodesic $\omega(t) = (1-t)x + ty$

Claim [AGS]:

1. W_2 is a geodesic space with geodesic between μ_0 and μ_1 given by the law of

(optimally coupled)

$$(1-t)X_0 + tX_1, \quad X_0 \sim \mu_0, \ X_1 \sim \mu_1$$



2. $\mathrm{curv}(W_2) \geq 0$

3. Statistical rates

- Ahidar-Coutrix, Le Gouic, Paris ('18)

$$\mathbb{E}\, W_2^2(\hat{b}, b^*) \lesssim n^{-1/2}$$

(Also: first rates on general Alexandrov Spaces)

- Le Gouic, Paris, R., Stromme ('19) [LPRS]

if $T_{b^* \to \mu}$ is gradient of $\begin{bmatrix}\beta\text{-smooth} \\ \alpha\text{-str. convex}\end{bmatrix}$ φ, $\forall\, \mu \in \mathrm{sub}(P)$, $\beta - \alpha < 1$

then

$$\mathbb{E}\, W_2^2(\hat{b}, b^*) \lesssim \frac{\sigma^2}{n}$$

where $\sigma^2 = \int W_2^2(b, b^*)\, dP(b)$

(+ similar results for general Alexandrov spaces)

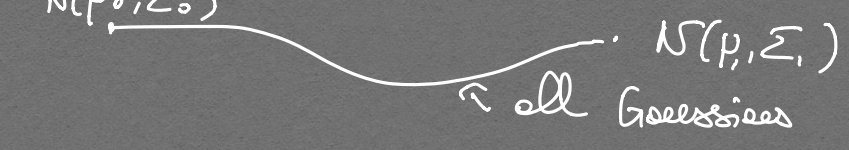
- Kroshnin, Spokoiny, Suvorikova

$$\mathbb{E}\, W_2^2(\hat{b}, b^*) \lesssim \frac{\sigma^2}{n}$$

if P is supported on Gaussians (+ CLT)

Remarks: the Gaussian case is interesting.

- "totally geodesic"



- Viable alternative to naive discretization.

[Cuturi & Doucet '14] (First order algorithm to compute Wasserstein barycenters)

$\mu_1, \dots, \mu_n \rightarrow$ discretize $\rightarrow$ algorithm.

↳ Statistical modeling → summarize a distribution in a few key parameters (e.g. mean, variance)

4- Gradient descent over the Wasserstein space

Want to minimize:

$$b \mapsto F(b) = \frac{1}{2} \int W_2^2(b, \mu)\, dP(\mu) \quad \text{over } \mathcal{W}_2$$

(e.g when $P = P_n = \frac{1}{n}\sum \delta_{\mu_i}$)

→ gradient descent.

Otto calculus: (see Ambrosio, Gigli, Savaré '08)

$f: \mathbb{R}^d \to \mathbb{R}$ $\qquad D_v f(x) = \lim_{h \to 0} \frac{f(x+hv) - f(x)}{h}$

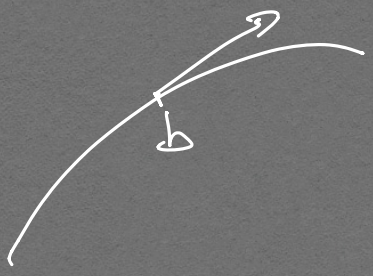
$x \in \mathbb{R}^d$

$v \in T_x \mathbb{R}^d = \mathbb{R}^d$

gradient $\nabla f(x) \in (T_x \mathbb{R}^d)^* = \mathbb{R}^d$ is s.t.

$$\langle \nabla f(x), v \rangle = D_v f(x) \quad \forall v \in T_x \mathbb{R}^d$$

Tangent space of $\mathcal{W}_2$ at b?



$T_{b \to \mu} - id$ (displacement map)
$= \nabla \varphi$, φ convex
is a tangent vector.

$$T_b \mathcal{W}_2 = \overline{\{ \lambda (\nabla \varphi - id), \lambda > 0, \varphi \text{ convex} \}}^{L^2}$$

↑ this is a subspace of $L^2(\mathbb{R}^d)$

$$\forall U, V \in T_b \mathcal{W}_2, \quad \langle V, U \rangle_b = \mathbb{E}_b \langle V(x), U(x) \rangle_{\mathbb{R}^d}$$

$$\exp_b : T_b \mathcal{W}_2 \to \mathcal{W}_2$$
$$V \longmapsto (id + V) \# b$$

Gradient of the barycenter functional $b \mapsto F(b)$

$V \in T_b \mathcal{W}_2$, $V = T_{b \to b'} - id$

$$D_V F(b) = \lim_{h \to 0} \frac{F(id + h(T_{b \to b'} - id) \# b) - F(b)}{h}$$

$$= \lim_{h\to 0} \frac{1}{2h} \int \Big[W_2^2\big(h(T_{b\to b'} - id)_{\#} b, \mu\big) - W_2^2(b,\mu) \Big] d\mathbb{P}(\mu)$$

$\hookrightarrow$ opt. btwn (b,μ)

$$\leqslant \int \| X + h(T_{b\to b'}(X) - X) - Y \|^2 \, d\gamma(X,Y)$$

$$= \int \|X-Y\|^2 + 2h \langle T_{b\to b'}(X) - X, X - Y \rangle d\gamma + O(h^2)$$

$$= W_2^2(X,Y) + 2h \, \langle T_{b\to b'} - id, id - T_{b\to\mu} \rangle_{L^2(b)} + O(h^2)$$

$$D_V F(b) \leqslant \int \langle T_{b\to b'} - id, id - T_{b\to\mu} \rangle_{L^2(b)} d\mathbb{P}(\mu)$$

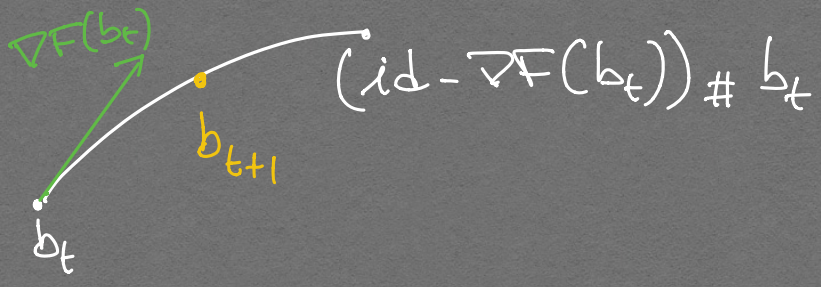
$$= \Big\langle \underbrace{T_{b\to b'} - id}_{V}, \underbrace{id - \int T_{b\to\mu} d\mathbb{P}_\mu}_{\nabla F(b)} \Big\rangle_b$$

$$\boxed{\nabla F(b) = id - \int T_{b\to\mu} \, d\mathbb{P}(\mu)}$$

Gradient descent:

$$b_{t+1} = \big[id - \eta_t \nabla F(b_t) \big]_{\#} b_t$$

$$= \Big[(1-\eta_t) id + \eta_t \underbrace{\int T_{b_t\to\mu} \, d\mathbb{P}(\mu)}_{\frac{1}{n}\sum_{i=1}^{n} T_{b_t\to\mu_i}} \Big]_{\#} b_t$$



SGD: $\rho_1, \dots, \rho_n \overset{iid}{\sim} P$

$$b_{t+1} = [(1-\eta_t)\, id + \eta_t\, T_{b_t \to \rho_t}]_\# b_t$$

Rates of convergence?

5- Optimization redux

$f: \mathbb{R}^d \to \mathbb{R}$ convex

$$\frac{\alpha}{2}\|x-y\|^2 \le f(y) - f(x) - \langle \nabla f(x), y-x\rangle \le \frac{\beta}{2}\|x-y\|^2$$

α - str. cvx

β - strongly smooth

GD: $x_{t+1} = x_t - \eta_t \nabla f(x_t)$

x^*: minimizer

$$f(x_{t+1}) - f(x^*) = f(x_t - \eta_t \nabla f(x_t)) - f(x^*)$$

$$\le f(x_t) - f(x^*) - \eta_t \|\nabla f(x_t)\|^2 + \frac{\eta_t^2 \beta}{2}\|\nabla f(x_t)\|^2$$

$\eta = \frac{1}{\beta}$

$$= f(x_t) - f(x^*) - \frac{1}{2\beta}\|\nabla f(x_t)\|^2$$

ASSUME Polyak-Łojasiewicz (PL) inequality:

$$\boxed{f(x) - f(x^*) \leq C_{PL} \|\nabla f(x)\|^2}$$

$$\leq \left(1 - \frac{1}{2\beta C_{PL}}\right)\left(f(x_t) - f(x^*)\right)$$

$$\leq \left(1 - \frac{1}{2\beta C_{PL}}\right)^{t+1} \left[f(x_0) - f(x^*)\right]$$

if α- strongly convex (at x^*), we get

$$\|x_t - x^*\|^2 \leq \frac{2}{\alpha}\left(1 - \frac{1}{2\beta C_{PL}}\right)^{t+1} \left[f(x_0) - f(x^*)\right]$$

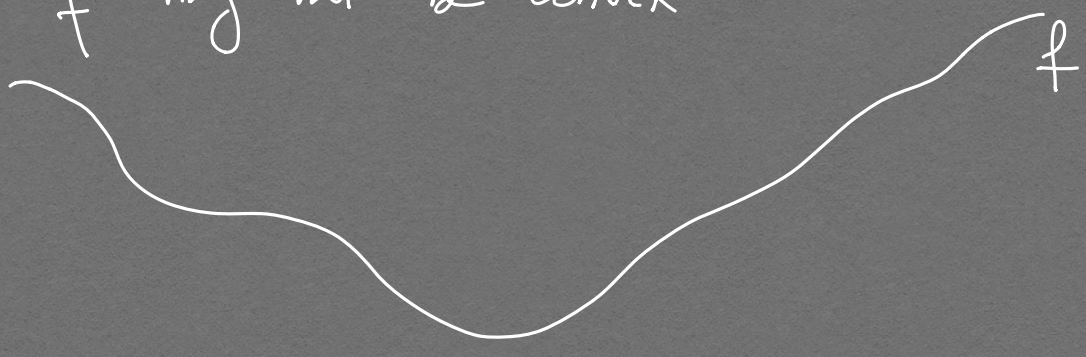
Recap: we used:

A. $f(y) - f(x) \leq \langle \nabla f(x), y - x \rangle + \frac{\beta}{2}\|y - x\|^2$

B. $\frac{\alpha}{2}\|x - x^*\|^2 \leq f(x) - f(x^*)$ at x^*

C. $f(x) - f(x^*) \leq C_{PL}\|\nabla f(x)\|^2$

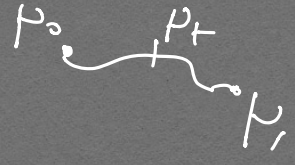
↳ Note: f may not be convex



6- Back to Wasserstein

$G: \mathcal{W}_2 \to \mathbb{R}$ is geodesically convex if $\forall$ geodesic $\mu_t \in \mathcal{W}_2$ from μ_0 to μ_1, it holds

$$G(\mu_t) \leq (1-t)G(\mu_0) + t\,G(\mu_1)$$



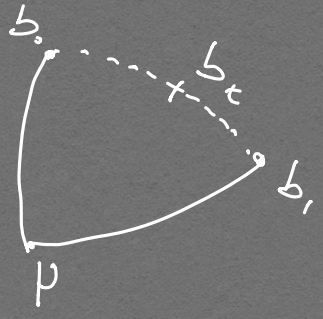
(can also define smoothness, strong convexity).

Example: if π is log-concave, then,

$$b \mapsto KL(b \,\|\, \pi) \text{ is g-convex.}$$

Fact: F is NOT geodesically convex

But: A. F is 1-smooth



$\text{Curv}(\mathcal{W}_2) \geq 0$

$$\int \frac{W_2^2(b_t,\mu) - W_2^2(b_0,\mu)}{2t}\, dP(\mu)$$

$$\geq \int (1-t)\, W_2^2(b_0,\mu) + t\, W_2^2(b_1,\mu) - \frac{t(1-t)}{2} W_2^2(b_0,b_1)\, dP(\mu)$$

$$= F(b_1) - F(b_0) - \frac{(1-t)}{2} W_2^2(b_0,b_1)$$

$t \to 0$:

$$\langle \nabla F(b_0), T_{b_0 \to b_1} - id \rangle_b \geq F(b_1) - F(b_0) - \frac{1}{2} W_2^2(b_0, b_1)$$

Rearranging:

$$F(b_1) - F(b_0) \leq \langle \nabla F(b_0), T_{b_0 \to b_1} - id \rangle_b + \frac{1}{2} W_2^2(b_0, b_1)$$

In fact 1-smoothness of d^2 is a characterization of $Curv \geq 0$.

B. $\quad F(b) - F(b^*) \geq C\, W_2^2(b, b^*)$

↳ Variance inequality.

Sturm ('03):

$$Curv(\mathcal{U}) \leq 0 \iff \int d^2(b,x) - d^2(b^*,x)\, dP(x) \geq 1 \cdot d^2(b,b^*) \quad \forall P$$

Th CMRS

Assume that $T_{b^* \to \mu} = \nabla \varphi_\mu$ for some $\alpha(\mu)$ str. cvx potential φ_μ

$$\Rightarrow \quad F(b) - F(b^*) \geq C_{var}\, W_2^2(b, b^*)$$

with $\quad C_{var} = \int \alpha(\mu)\, dP(\mu)$

Rk: strict improvement over LPRS (no smoothness)

C. Integrated PL inequality

Kantorovich duality yields:

$$\frac{1}{2} W_2^2(b,\mu) = \int \left(\frac{\|\cdot\|^2}{2} - \varphi_{\mu\to b} \right) d\mu + \int \left(\frac{\|\cdot\|^2}{2} - \varphi_{b\to\mu} \right) db$$

and, $\forall b'$, $\quad \frac{1}{2} W_2^2(b^*,\mu) \geq \int \left(\frac{\|\cdot\|^2}{2} - \varphi_{\mu\to b} \right) d\mu + \int \left(\frac{\|x\|^2}{2} - \varphi_{b\to\mu} \right) db^*$

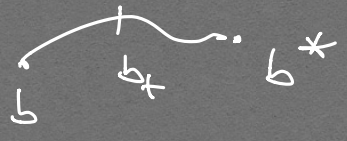
$$\Rightarrow \quad F(b) - F(b^*) \leq \int \underbrace{\left(\frac{\|\cdot\|^2}{2} - \int \varphi_{\mu\to b} \, dP(\mu) \right)}_{f} d(b - b^*)$$

take $X \sim b$, $Y \sim b^*$ optimally coupled

$$\mathbb{E}[f(X) - f(Y)] = \int_0^1 \mathbb{E}[\langle \nabla f(tY + (1-t)X), Y - X \rangle] dt$$

$$\leq \sqrt{\int_0^1 \mathbb{E} \|\nabla f(X_t)\|^2 dt} \sqrt{\mathbb{E}\|Y - X\|^2}$$

$$= W_2(b, b^*) \left(\int_0^1 \|\nabla f\|^2_{L^2(b_t)} dt \right)^{1/2}$$



$$\leq \sqrt{\frac{F(b) - F(b^*)}{C_{VAR}}} \left(\int_0^1 \|\nabla f\|^2_{L^2(b_t)} dt \right)^{1/2}$$

$$\Rightarrow \quad \boxed{F(b) - F(b^*) \leq \int_0^1 \|\nabla f\|^2_{L^2(b_t)} dt}$$

to obtain a real PL, need to control

$$\|\nabla f\|^2_{L^2(b_t)} \le \|\nabla f\|^2_{L^2(b)} \qquad \forall\, t \in [0,1]$$

Would be sufficient to have $\frac{db_t}{db} \le C$

Main issue: No control on iterates of GD.

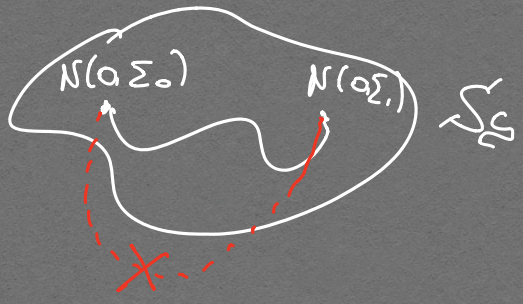
7. Gaussian distributions

$$\mathrm{supp}(P) \subset S_\xi$$

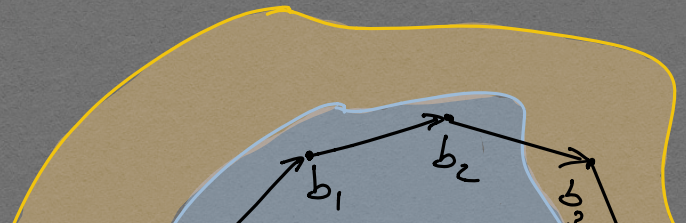
$$S_\xi := \{ \mathcal{N}(0,\Sigma) : \|\Sigma\|_{op} \le 1,\ \det \Sigma \ge 2\xi \}$$

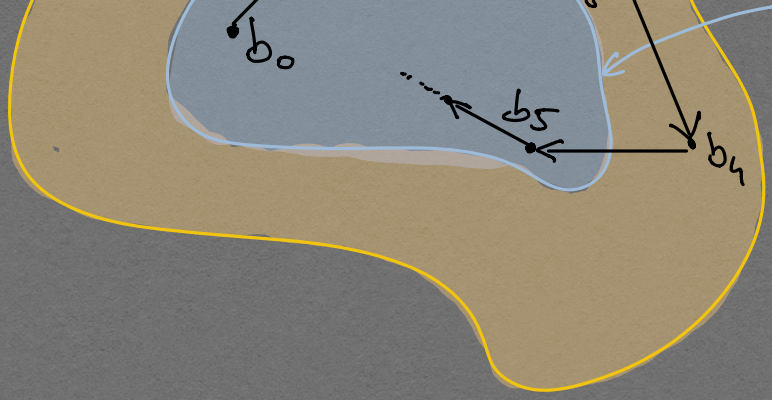
density upper bound.

Claim: S_ξ is a geodesically convex set



Recall: GD & SGD push along geodesics





A. (1-Smoothness) is always true

B. Variance inequality :

$$F(b) - F(b^*) \geq \frac{\xi}{2} W_2^2(b, b^*)$$

C.

Th] CMRS 20

$\forall b \in \mathcal{S}_\xi$:

$$F(b) - F(b^*) \leq \frac{1}{2\xi^2} \|\nabla F(b)\|^2_{L^2(b)}$$

(PL ineq

Concretely

Algorithm 1 Bures-Wasserstein GD

procedure BURES-GD(Σ_0, Q, T)
  for $t = 1, \dots, T$ **do**
    $S_t \leftarrow \int \Sigma_{t-1}^{-1/2} \{\Sigma_{t-1}^{1/2} \Sigma(\mu) \Sigma_{t-1}^{1/2}\}^{1/2} \Sigma_{t-1}^{-1/2} \, \mathrm{d}Q(\mu)$
    $\Sigma_t \leftarrow S_t \Sigma_{t-1} S_t$
  end for
  return Σ_T
end procedure

Algorithm 2 Bures-Wasserstein SGD

procedure BURES-SGD($\Sigma_0, (\eta_t)_{t=1}^T, (K_t)_{t=1}^T$)
  for $t = 1, \dots, T$ **do**
    $\hat{S}_t \leftarrow \Sigma_{t-1}^{-1/2} \{\Sigma_{t-1}^{1/2} K_t \Sigma_{t-1}^{1/2}\}^{1/2} \Sigma_{t-1}^{-1/2}$
    $\Sigma_t \leftarrow ((1-\eta_t) I_D + \eta_t \hat{S}_t) \Sigma_{t-1} ((1-\eta_t) I_D + \eta_t \hat{S}_t)$
  end for
  return Σ_T
end procedure

↑ Studied in Alvarez-Esteban, del Barrio & Cuesta-Albertos '16 as fixed point algorithm ($\nabla F(b^*) = 0$. No guarantees)

Main results

Run either GD on $F(b) = \frac{1}{n}\sum_{i=1}^{n} W_2^2(b, \mu_i)$

or SGD on $\mu_1, \dots, \mu_n$

Th GD with stepsize 1

$$W_2^2(b_T, \hat{b}_n) \lesssim \frac{(1-\Sigma^2)^T}{\Sigma}$$

$\Downarrow$ + Kroshnin et al.

$$\mathbb{E}\, W_2^2(b_T, b^*) \lesssim n^{-1/2}$$

for $T \geq C \log n$

Th SGD with $\eta_t \sim \frac{1}{t}$

$$\mathbb{E}\, W_2^2(b_n, b^*) \lesssim \frac{\mathrm{Var}(P)}{n\,\Sigma^5}$$

Some open questions

1. Better dependence on Σ ?
2. Averaging
3. Other fuctionals ?
4. Beyond Gaussians.