Learning and Optimization for Convex Problems

Learning *using* Optimization

- **Goal of (supervised) learning**: find predictor $h_w: \mathcal{X} \to \mathcal{Y}$ with low expected error

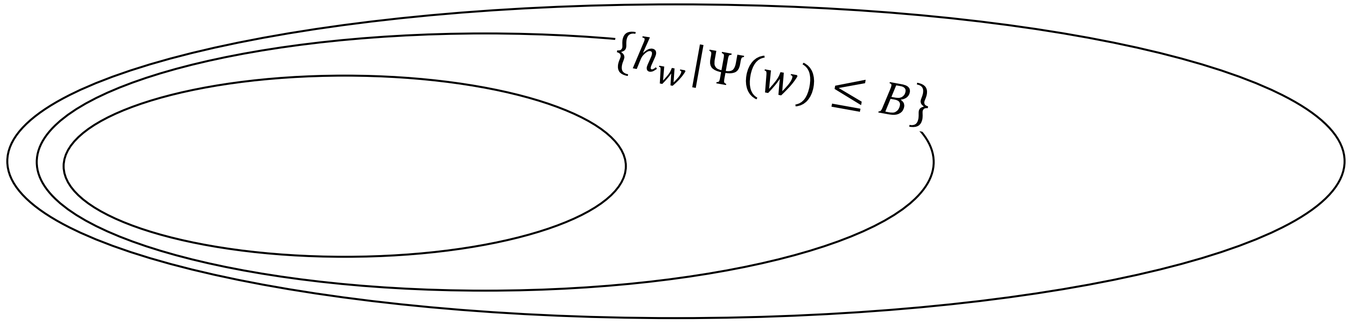
$$L(h_w) = \mathbb{E}_{x,y}[loss(h_w(x); y)]$$

- Collect $S = \{(x_1, y_1) \ldots (x_m, y_m)\}$ and minimize empirical objective:

$$\hat{h}_w = \arg\min_{h \in \mathcal{H}} \frac{1}{m} \sum_i loss(h_w(x_i); y_i)$$

or

$$\hat{h}_w = \arg\min_{\Psi(w) \leq B} \frac{1}{m} \sum_i loss(h_w(x_i); y_i)$$



Generalization: Uniform convergence in $\mathcal{H}$ or $\{h_w | \Psi(w) \leq B\}$

➔ $|L(h_w) - \hat{L}(h_w)|$ small ➔ hence $L(h_w)$ low

Example: $\mathcal{H} = \{h_w(x) \mapsto \langle w, x\rangle \mid \|w\|_2 \leq B\}, \|x\|_2 \leq 1, \quad |\ell'| \leq 1$

$$\hat{w} = \arg\min_{\{\|w\|\leq B\}} \hat{L}(w) \qquad L(\hat{w}) \leq \inf_{\|w\|\leq B} L(w) + O\left(\sqrt{\frac{B^2}{m}}\right)$$

Gradient Descent on $\hat{L}(w) = \frac{1}{m}\sum_i \ell(\langle w, x_i\rangle, y_i)$:

$$w^{(k+1)} = w^{(k)} - \eta\nabla\hat{L}\left(w^{(k)}\right)$$

Optimization Guarantee: $\hat{L}\left(\bar{w}^{(T)}\right) \leq \inf_{\|w\|\leq B} \hat{L}(w) + O\left(\sqrt{\frac{B^2}{T}}\right)$

$\Rightarrow L\left(\bar{w}^{(T)}\right) \leq \inf_{\|w\|\leq B} L(w) + O\left(\sqrt{\frac{B^2}{m}}\right) + O\left(\sqrt{\frac{B^2}{T}}\right)$

Example: $\mathcal{H} = \{h_w(x) \mapsto \langle w, x\rangle \mid \|w\|_2 \le B\}, \|x\|_2 \le 1, \ |\ell'| \le 1$

$$\hat{w} = \arg\min_{\{\|w\| \le B\}} \hat{L}(w) \qquad L(\hat{w}) \le \inf_{\|w\| \le B} L(w) + O\left(\sqrt{\frac{B^2}{m}}\right)$$

- Stochastic Gradient Descent on $\hat{L}(w) = \frac{1}{m}\sum_i \ell(\langle w, x_i\rangle, y_i)$:

$$w^{(k+1)} = w^{(k)} - \eta \nabla \ell\left(\left\langle w^{(k)}, x_i\right\rangle, y_i\right)$$

$$\mathbb{E}_i\left[\ell\left(\left\langle w^{(k)}, x_i\right\rangle, y_i\right)\right] = \nabla \hat{L}\left(w^{(k)}\right)$$

Optimization Guarantee: $\hat{L}\left(\bar{w}^{(T)}\right) \le \inf_{\|w\| \le B} \hat{L}(w) + O\left(\sqrt{\frac{B^2}{T}}\right)$

$\Rightarrow L\left(\bar{w}^{(T)}\right) \le \inf_{\|w\| \le B} L(w) + O\left(\sqrt{\frac{B^2}{m}}\right) + O\left(\sqrt{\frac{B^2}{T}}\right)$

- One-Pass SGD viewed as SGD on $L(w) = \mathbb{E}[\ell(\langle w, x\rangle, y)]$:

$$\mathbb{E}_{x_i, y_i}\left[\ell\left(\left\langle w^{(k)}, x_i\right\rangle, y_i\right)\right] = \nabla L\left(w^{(k)}\right)$$

Optimization Guarantee: $L\left(\bar{w}^{(T)}\right) \le \inf_{\|w\| \le B} L(w) + O\left(\sqrt{\frac{B^2}{T}}\right)$

One-pass: #itter T = #samples m

Learning *is* Stochastic Optimization

$$\min_{x \in \mathcal{X}} F(x) = \mathbb{E}_{z \sim \mathcal{D}}[f(x,z)]$$

based on i.i.d samples $z_1, z_2, z_3, \ldots \sim \mathcal{D}$

- Distribution $\mathcal{D}$ unknown; No direct access to $F(x)$
- Can obtain unbiased estimates of $F(x)$, $\nabla F(x)$, etc

- Learning as stochastic optimization:

$$\min_{h:\mathcal{X}\to\mathcal{Y}} L(h) = E_{x,y\sim\mathcal{D}}[\underbrace{loss(h(x),y)}_{f(h,(x,y)) = loss(h(x),y)}]$$

based on sample $(x_1, y_1), \ldots, (x_m, y_m) \sim \mathcal{D}$

- Vapnik's "General Learning Setting" ***is*** stochastic optimization:

$$\min_h L(h) = \mathbb{E}_z[\ell(h,z)]$$

based on sample $z_1, z_2, \ldots \sim \mathcal{D}$

Optimization	**Statistics**	**COLT**	**NeurIPS**
x	β	h	w

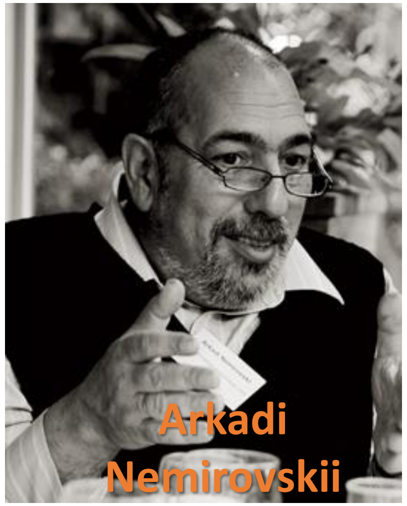
Stochastic Optimization ≡ General Learning

$$\min_{h\in\mathcal{H}} F(h) = \mathbb{E}_{z\sim\mathcal{D}}[f(h,z)] \text{ based on } z_1,\dots,z_m \sim iid\ \mathcal{D}$$

- Supervised learning:
  - $\mathcal{Z} = \mathcal{X}\times\mathcal{Y} = \{z = (x,y) \mid x\in\mathcal{X}, y\in\mathcal{Y}\}$
  - $h: \mathcal{X}\to\mathcal{Y}$
  - $f(h,z) = loss(h(x); y)$
- Unsupervised learning, e.g. k-means clustering:
  - $z = x \in \mathbb{R}^d$,
  - $h = (\mu[1],\mu[2],\dots,\mu[k]) \in \mathbb{R}^{d\times k}$ specified k cluster centers
  - $f\big((\mu[1],\mu[2],\dots,\mu[k]), x\big) = \min_i \|\mu[i]-x\|^2$
- Density estimation:
  - $z = x$ in some measurable space $\mathcal{Z}$ (e.g. $\mathbb{R}^d$)
  - h specifies probability density $p_h(z)$
  - $f(h,z) = -\log p_h(z)$
- Learning a good route with random traffic:
  - $z =$ traffic delays on each road segment
  - $h =$ route chosen (indicator over road segments in route)
  - $f(h,z) = \langle h,z\rangle =$total delay along route

Stochastic Optimization vs Statistical Learning

- Focus on computational efficiency
- Generally assumes unlimited sampling
  - as in monte-carlo methods for complicated objectives
- Optimization variable generally a vector in a normed space
  - complexity control through norm
- Mostly convex objectives



- Focus on sample size
- What can be done with a fixed number of samples?
- Abstract hypothesis classes
  - linear predictors, but also combinatorial hypothesis classes
  - generic measures of complexity such as VC-dim, fat shattering, Radamacher
- Also non-convex classes and loss functions

Stochastic Optimization (≡ Learning)

$$\min_{w \in \mathcal{W}} F(w) = \mathbb{E}_{z \sim \mathcal{D}}[f(w, z)]$$

based on i.i.d samples $z_1, z_2, z_3, \ldots \sim \mathcal{D}$

- Sample Average Approximation (SAA)/Empirical Risk Minimization (ERM):
  - Collect sample $z_1, \ldots, z_m$
  - Minimize $\hat{F}_m(w) = \frac{1}{m}\sum_i f(w, z_i)$
- Stochastic Approximation (SA), e.g. Stochastic Gradient Descent (SGD):
  - Update $w^{(i)}$ based on $f\left(w^{(i)}, z_i\right), \nabla f\left(w^{(i)}, z_i\right)$, etc
  - E.g. $w^{(i+1)} = w^{(i)} - \eta \nabla f\left(w^{(i)}, z_i\right)$

SGD for Machine Learning

$$\min_{w} L(w)$$

Direct SA Approach:

Initialize $w^{(0)} = 0$

At iteration t:

- Draw $x_t, y_t \sim \mathcal{D}$
- $w^{(t+1)} \leftarrow w^{(t)} - \eta_t \nabla \ell\left(\langle w^{(t)}, x_t \rangle, y_t\right)$

Return $\bar{w}^{(T)} = \frac{1}{T}\sum_{t=1}^{T} w^{(t)}$

- Fresh sample at each iteration, $m = T$
- No need to project nor require $\|w\| \leq B$
- Implicit regularization via early stopping

SGD on ERM:

$$\min_{\|w\|_2 \leq B} L_S(w)$$

Draw $(x_1, y_1), \ldots, (x_m, y_m) \sim \mathcal{D}$

Initialize $w^{(0)} = 0$

At iteration t:

- Pick $i \in 1 \ldots m$ at random
- $w^{(t+1)} \leftarrow w^{(t)} - \eta_t \nabla \ell\left(\langle w^{(t)}, x_i \rangle, y_i\right)$
- $w^{(t+1)} \leftarrow proj\ w^{(t+1)}\ to\ \|w\| \leq B$

Return $\bar{w}^{(T)} = \frac{1}{T}\sum_{t=1}^{T} w^{(t)}$

- Can have $T > m$ iterations
- Need to project to $\|w\| \leq B$
- Explicit regularization via $\|w\|$

SGD for Machine Learning

$$\min_w L(w)$$

Direct SA Approach:

Initialize $w^{(0)} = 0$

At iteration t:

- Draw $x_t, y_t \sim \mathcal{D}$
- $w^{(t+1)} \leftarrow w^{(t)} - \eta_t \nabla \ell(\langle w^{(t)}, x_t \rangle, y_t)$

$\eta_t = \sqrt{B^2 t}$

Return $\bar{w}^{(T)} = \frac{1}{T} \sum_{t=1}^{T} w^{(t)}$

$$L(\bar{w}^{(T)}) \le L(w^*) + \sqrt{\frac{B^2}{T}}$$

SGD on ERM:

$$\min_{\|w\|_2 \le B} L_S(w)$$

Draw $(x_1, y_1), \dots, (x_m, y_m) \sim \mathcal{D}$

Initialize $w^{(0)} = 0$

At iteration t:

- Pick $i \in 1 \dots m$ at random
- $w^{(t+1)} \leftarrow w^{(t)} - \eta_t \nabla \ell(\langle w^{(t)}, x_i \rangle, y_i)$
- $w^{(t+1)} \leftarrow proj\ w^{(t+1)}\ to\ \|w\| \le B$

Return $\bar{w}^{(T)} = \frac{1}{T} \sum_{t=1}^{T} w^{(t)}$

$$L(\bar{w}^{(T)}) \le L(w^*) + 2\sqrt{\frac{B^2}{m}} + \sqrt{\frac{B^2}{T}}$$

$w^* = \arg\min_{\|w\| \le B} L(w)$

SGD for Machine Learning

$$\min_w L(w)$$

Direct SA Approach:

Initialize $w^{(0)} = 0$
At iteration t:
- Draw $x_t, y_t \sim \mathcal{D}$
- $w^{(t+1)} \leftarrow w^{(t)} - \eta_t \nabla \ell(\langle w^{(t)}, x_t \rangle, y_t)$

Return $\bar{w}^{(T)} = \frac{1}{T} \sum_{t=1}^{T} w^{(t)}$

- Fresh sample at each iteration, $m = T$
- No need to project nor require $\|w\| \leq B$
- Implicit regularization via early stopping

SGD on RERM:

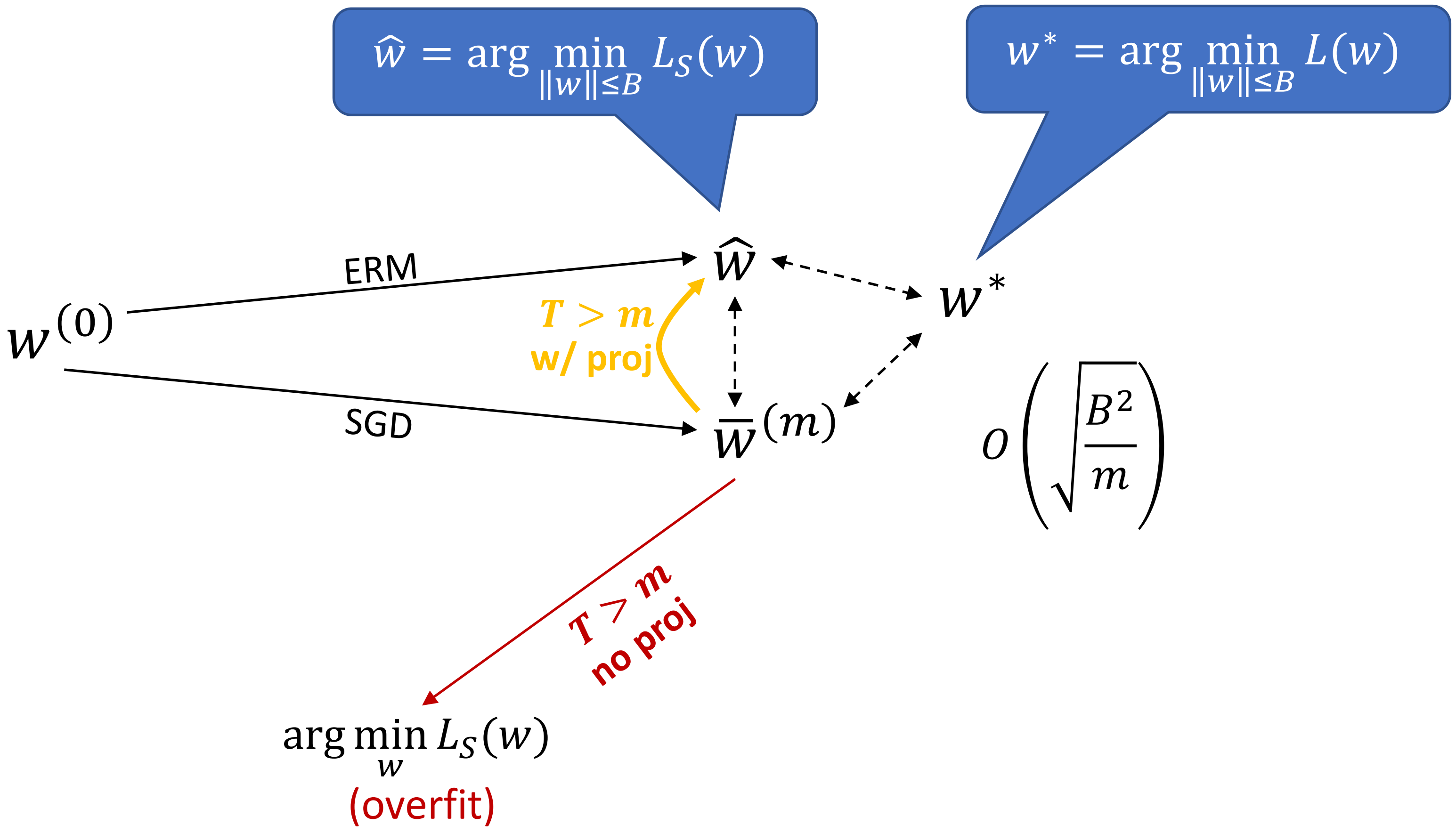
$$\min L_S(w) + \frac{\lambda}{2}\|w\|$$

Draw $(x_1, y_1), \ldots, (x_m, y_m) \sim \mathcal{D}$
Initialize $w^{(0)} = 0$
At iteration t:
- Pick $i \in 1 \ldots m$ at random
- $w^{(t+1)} \leftarrow w^{(t)} - \eta_t \nabla \ell(\langle w^{(t)}, x_i \rangle, y_i) - \boldsymbol{\lambda w}$

Return $\bar{w}^{(T)} = \frac{1}{T} \sum_{t=1}^{T} w^{(t)}$

- Can have $T > m$ iterations
- Need to shrink w
- Explicit regularization via $\|w\|$

SGD vs ERM

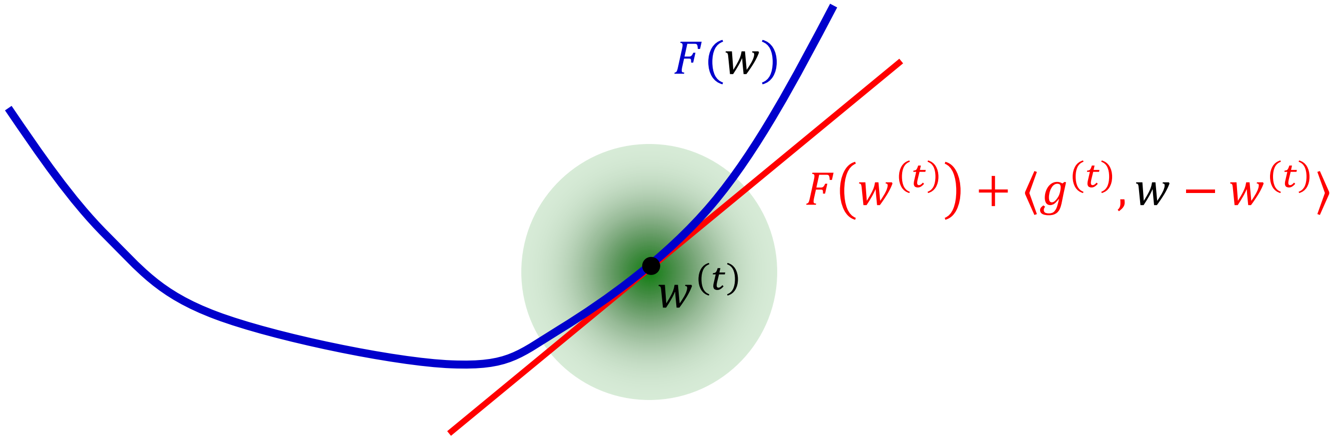


$\hat{w} = \arg\min_{\|w\| \le B} L_S(w)$
$w^* = \arg\min_{\|w\| \le B} L(w)$
ERM
$w^{(0)}$
$\hat{w}$
w^*
$T > m$
w/ proj
SGD
$\bar{w}^{(m)}$
$O\left(\sqrt{\frac{B^2}{m}}\right)$
$T > m$
no proj
$\arg\min_{w} L_S(w)$
(overfit)

Where's the Regularization

- Gradient Descent seems to be regularizing with $\|w\|_2$. How?

$$w^{(t+1)} \leftarrow \arg\min_w \underbrace{F\left(w^{(t)}\right) + \langle g^{(t)}, w - w^{(t)} \rangle}_{\text{1st order model of F}(\mathbf{w}) \text{ around } \mathbf{w}^{(t)}\text{, based on } \mathbf{g}^{(t)}} + \underbrace{\frac{1}{2\eta}\left\|w - w^{(t)}\right\|_2}_{\text{only valid near w}^{(t)}\text{, so don't go too far}}$$



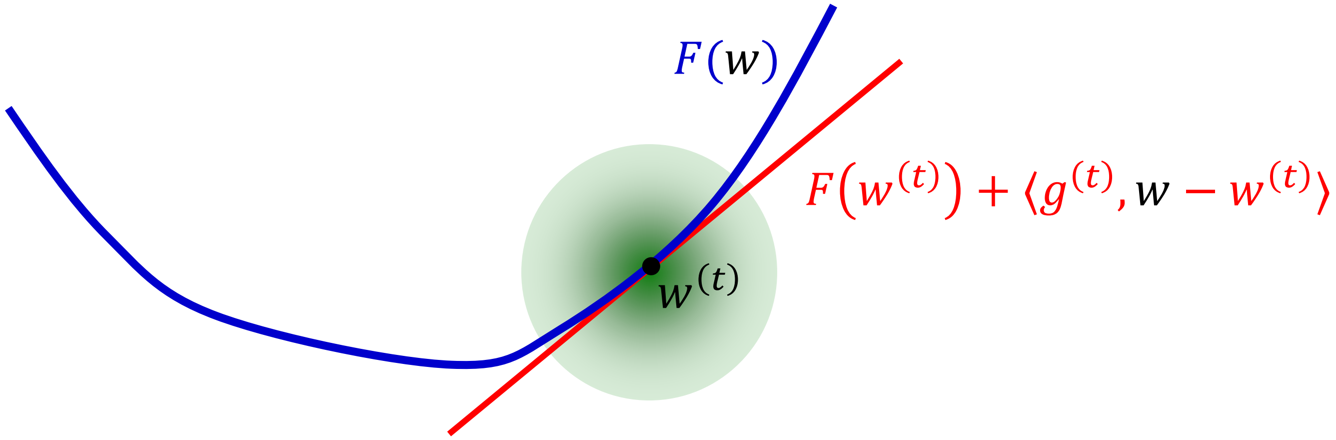
Where's the Regularization

- Gradient Descent seems to be regularizing with $\|w\|_2$. How?

$$w^{(t+1)} \leftarrow \arg\min_{w} \underbrace{F\left(w^{(t)}\right) + \langle g^{(t)}, w - w^{(t)} \rangle} + \frac{1}{2\eta}\left\|w - w^{(t)}\right\|_2$$

$$= \arg\min_{w} \langle g^{(t)}, w \rangle + \frac{1}{2\eta}\left\|w - w^{(t)}\right\|_2$$

$$= w^{(t)} - \eta g^{(t)}$$



Stability

- Definition: learning rule $\widetilde{\mathrm{w}}(z_1, \dots z_m)$ is (leave-one-out) $\beta(m)$-*stable* if:
$$|f(\widetilde{w}(z_1, \dots, z_{m-1}), z_m) - f(\widetilde{w}(z_1, \dots, z_m), z_m)| \le \beta(m)$$

- Theorem: If $\widetilde{w}$ is symmetric and $\beta(m)$-stable ➔
$$\mathbb{E}[F(\widetilde{w}_{m-1})] \le \mathbb{E}\left[\hat{F}(\widetilde{w}_m)\right] + \beta(m)$$

Proof of Theorem:

$$\begin{aligned}
\mathbb{E}_{z_1,\dots,z_{m-1}\sim\mathcal{D}}[F(\widetilde{w})] &= \mathbb{E}_{z_1,\dots,z_m}[f(\widetilde{w}(z_1, \dots, z_{m-1}), z_m)] \\
&= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}[f(\widetilde{w}(z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_m), z_i)] \\
&\le \frac{1}{m}\sum_{i=1}^{m}\left(\mathbb{E}[f(\widetilde{w}(z_1, \dots, z_m), z_i)] + \beta(m)\right) \\
&= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} f(\widetilde{w}(z_1, \dots, z_m), z_i)\right] + \beta(m) = \mathbb{E}\left[\hat{F}(\widetilde{w}_m)\right] + \beta(m)
\end{aligned}$$

Strong Convexity

- **Definition**: $\Psi: \mathcal{W} \to \mathbb{R}$ is α-strongly convex w.r.t a norm $\|w\|$ if

$$\forall_{w,w' \in \mathcal{W}} \Psi(w') \geq \Psi(w) + \langle \nabla\Psi(w), w' - w \rangle + \frac{\alpha}{2}\|w' - w\|^2$$

- E.g. $\Psi(w) = \frac{1}{2}\|w\|_2^2$ is 1-strongly convex w.r.t $\|w\|_2$

Proof: $\frac{1}{2}\|w\|_2^2 + \langle w, w' - w \rangle + \frac{1}{2}\|w' - w\|_2^2 = \|w + (w' - w)\|_2^2 = \|w'\|^2$

- **Claim**: if Ψ is α-strongly convex, and $w_0 = \arg\min_{w \in \mathcal{W}} \Psi(w)$, then

$$\forall_{w \in \mathcal{W}} \Psi(w) - \Psi(w_0) \geq \frac{\alpha}{2}\|w - w_0\|^2$$

- **Claim**: if Ψ is α-strongly convex, then cΨ is $(c \cdot \alpha)$-strongly convex
- **Claim**: if $f(w)$ is convex and $\Psi(w)$ is α-strongly convex, then $f(w) + \Psi(w)$ is α-strongly convex

- **Definition**: $\Psi(w)$ is α-s.c. w.r.t $\|w\|$ if $\forall_{w,w' \in \mathcal{W}} \Psi(w') \geq \Psi(w) + \langle \nabla\Psi(w), w' - w \rangle + \frac{\alpha}{2}\|w' - w\|^2$

$$\mathrm{RERM}_{\lambda\Psi}(S) = \arg\min_{w \in \mathcal{W}} F_S(w) + \lambda\Psi(w)$$

- **Definition**: $f(w,z)$ is G-Lipschitz w.r.t $\|w\|$ iff $\forall_{z \in \mathcal{Z}} \forall_{w,w' \in \mathcal{W}} |f(w,z) - f(w',z)| \leq G \cdot \|w' - w\|$
$(\equiv \|\nabla_w f(w,z)\|_* \leq G)$

- **Claim**: f is G-Lipschitz and $\Psi(w)$ is α-s.c. ➔ $\mathrm{RERM}_{\lambda\Psi}(S)$ then is $\beta(m) \leq \frac{2G^2}{m\lambda\alpha}$ stable

- **Learning with $\boldsymbol{RERM_{\lambda\Psi}}$:**

$$\mathbb{E}[L_{\mathcal{D}}(RERM_{\lambda\Psi}(S))] \leq \mathbb{E}[F_S(RERM_{\lambda\Psi}(S))] + \beta(m)$$

$$\leq \mathbb{E}[F_S(RERM_{\lambda\Psi}(S)) + \lambda\Psi(w)] + \beta(m)$$ for $\Psi \geq 0$

$$\leq \mathbb{E}[F_S(w) + \lambda\Psi(w)] + \beta(m) = L_{\mathcal{D}}(w) + \lambda\Psi(w) + \frac{2G^2}{\lambda\alpha m}$$

$$\leq \inf_{w \in \mathcal{H}} F(w) + \sqrt{\frac{8\left(\sup_{w \in \mathcal{W}} \Psi(w)\right) G^2}{\alpha m}}$$

$\lambda = \sqrt{2G^2 / \alpha m(\sup \Psi(w))}$

$$\min_{w\in\mathcal{W}} \mathbb{E}_{z\sim\mathcal{D}}[f(w,z)] \quad \text{over convex } \mathcal{W}$$

$$f\big(w,(x,y)\big) = loss(\langle w,\phi(x)\rangle; y)$$

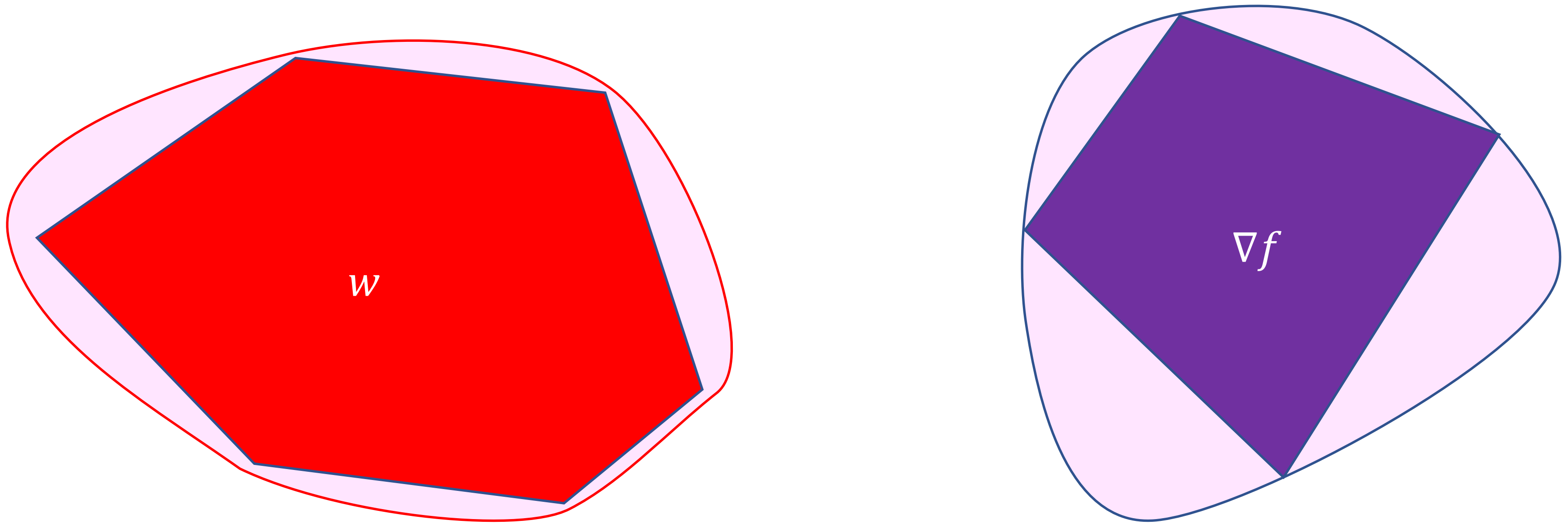
- The problem is **convex** if for every z, $f(w,z)$ is convex in w
  - If $loss(\hat{y};y)$ is convex in $\hat{y}$, the problem is convex
  - For a non-trivial loss, e.g. $loss(\hat{y},y) = |\hat{y}-y|$, $loss(h_w(x),y)$ is convex in w **only** when $h_w(x) = \langle w,\phi(x)\rangle$
- f is ***G*-Lipschitz with respect to a norm $\|w\|$** ($\|\nabla_w f(w,z)\|_* \le G$)
  - If $loss(\hat{y};y)$ is g-Lipschitz in $\hat{y}$ (as a scalar function): $\big|f\big(w,(x,y)\big) - f\big(w',(x,y)\big)\big| \le g\|\phi(x)\|_* \cdot \|w-w'\|$

    ➔ If $\|\phi(x)\|_* \le R$ for the dual norm, then the problem is $G = gR$ Lipschitz w.r.t $\|w\|$
- Bounded w.r.t some Ψ: $\Psi(w^*) \le B$
- $\Psi(w)$ is α-s.c. w.r.t $\|w\|$

$$\mathbb{E}[F(\hat{w}_{\lambda\Psi})] - F(w^*) \le O\left(\sqrt{\frac{\Psi(w^*)\,(\sup\|\nabla f\|_*)}{m}}\right)$$

$$➔\ m = O\big(\Psi(w^*)\,(\sup\|\nabla f\|_*)\big)$$

Matching the Geometry



$f\big(w,(x,y)\big) = \ell(\langle w, \phi(x)\rangle, y)$

$$\mathbb{E}[F(\hat{w}_{\lambda\Psi})] - F(w^*) \le O\left(\sqrt{\frac{\Psi(w^*)\,(\sup\|\nabla f\|_*)}{m}}\right) = O\left(\sqrt{\frac{\Psi(w^*)\,(\sup\|\phi(x)\|_*)}{m}}\right)$$

$$\Rightarrow m = O\big(\Psi(w^*)\,(\sup\|\nabla f\|_*)\big) = O\big(\Psi(w^*)\,(\sup\|\phi(x)\|_*)\big)$$

Matching the Geometry

- $\Psi(w) = \frac{1}{2}\|w\|_2^2$ is 1-strongly convex w.r.t $\|w\|_2$
$$m \propto \|w\|_2^2 \cdot \|x\|_2^2$$
- $\Psi(w) = \frac{1}{2}w^T Q w$ is 1-strongly convex w.r.t $\|w\|_Q = \sqrt{w^T Q w}$
$$m \propto (w^T Q w)(x^T Q^{-1} x)$$
- $\Psi(w) = \frac{1}{2(p-1)}\|w\|_p^2$ is 1-strongly convex w.r.t. $\|w\|_p$
$$m \propto \frac{\|w\|_p^2 \|x\|_q^2}{(p-1)}$$
- $\Psi(w) = \sum_i w[i] \log \frac{w[i]}{1/d}$ is 1-strongly convex w.r.t $\|w\|_1$ on $\{w > 0 \,|\, \|w\|_1 \leq 1\}$
$$m \propto \|w\|_1^2 \|x\|_\infty^2 \log(d)$$

$$\mathbb{E}[F(\widehat{w}_{\lambda\Psi})] - F(w^*) \leq O\left(\sqrt{\frac{\Psi(w^*)\,(\sup\|\nabla f\|_*)}{m}}\right) = O\left(\sqrt{\frac{\Psi(w^*)\,(\sup\|\phi(x)\|_*)}{m}}\right)$$

$$\Rightarrow m = O\big(\Psi(w^*)\,(\sup\|\nabla f\|_*)\big) = O\big(\Psi(w^*)\,(\sup\|\phi(x)\|_*)\big)$$

- RERM (offline, batch):

$$\hat{w} = \arg\min F_S(w) + \lambda \tfrac{1}{2}\|w\|_2$$

For general $\Psi(w)$:

$$\hat{w} = \arg\min F_S(w) + \lambda \Psi(w)$$

- Online / Stochastic Approximation:

SGD: $$w_{t+1} = \arg\min_{w} \langle \nabla f(w_t, z_t), w \rangle + \lambda_t \tfrac{1}{2}\|w - w_t\|_2$$

For general $\Psi(w)$???

Online Optimization (Learning)

$$
\begin{array}{lccccccc}
\text{Adversary:} & & f(w_1,z_1) & & f(w_2,z_2) & & f(w_3,z_3) & \\
& \nearrow & & \searrow\ \nearrow & & \searrow\ \nearrow & & \searrow \\
\text{Optimizer:} & w_1 & & w_2 & & w_3 & & \cdots
\end{array}
$$

- Arbitrary unknown sequence $z_1, z_2, \ldots \in \Omega$ (not stochastic/iid)
- Online learning rule: $w_i(z_1, \ldots, z_{i-1})$
- Goal: minimize Online Regret: for any sequence,

$$\frac{1}{m}\sum_{i=1}^{m} f(w_i(z_1,\ldots,z_{i-1}),z_i) \le \inf_{w\in\mathcal{W}} \frac{1}{m}\sum_{i=1}^{m} f(w,z_i) + Reg(m)$$

- Online $Reg(m)$ ➔ suboptimality of $\bar{w}_m = \frac{1}{m}\sum_i w_i$ for stochastic problem $F(w) = \mathbb{E}_z[f(w,z)]$

$$\mathbb{E}[F(\bar{w}_m)] \le \mathbb{E}\left[\frac{1}{m}\sum_i F(w_i)\right] = \mathbb{E}\left[\frac{1}{m}\sum_i f(w_i,z_i)\right] \le \mathbb{E}\left[\frac{1}{m}\sum_i f(w^*,z_i) + Reg(m)\right] = F(w^*) + Reg(m)$$

Stability in Online Learning

- Reminder: rule $\tilde{w}(z_1, \dots z_m)$ is $\beta(m)$-*stable* if
$$\left|f(\tilde{w}(z_1, \dots, z_{m-1}), z_m) - f(\tilde{w}(z_1, \dots, z_m), z_m)\right| \le \beta(m)$$
- Follow The Leader (FTL): $\hat{w}_m(z_1, \dots, z_{m-1}) = \arg\min_{w \in \mathcal{W}} \sum_{i=1}^{m-1} f(w, z_i)$
- Be The Leader (BTL) [a rule for prophets]: $w_m(z_1, \dots, z_{m-1}) = \arg\min_{w \in \mathcal{W}} \sum_{i=1}^{m} f(w, z_i)$
- If the ERM is $\beta(m)$-stable: $Reg_{FTL}(m) \le \underbrace{Reg_{BTL}(m)}_{\le 0} + \frac{1}{m}\sum_i \beta(i) \le \frac{1}{m}\sum_i \beta(i)$
- Follow The Regularized Leader (FTRL): $\hat{w}_m^{\lambda}(z_1, \dots, z_{m-1}) = \arg\min_{x} \sum_{i=1}^{m-1} f(w, z_i) + \lambda_i \Psi(w)$
- If f is convex and Lipschitz and Ψ is strongly conv. both w.r.t. $\|\cdot\|$:
$$Reg_{FTRL}(m) \le \sqrt{\frac{\Psi(w^*) \sup \|\nabla f\|}{m}}$$

- RERM (offline, batch):

$$\hat{w} = \arg\min F_S(w) + \lambda \tfrac{1}{2}\|w\|_2$$

For general $\Psi(w)$:

$$\hat{w} = \arg\min F_S(w) + \lambda\Psi(w)$$

- Online / Stochastic Approximation:

SGD:

$$w_{t+1} = \arg\min_{\mathrm{w}} \langle \nabla f(w_t, z_t), w\rangle + \lambda_t \tfrac{1}{2}\|w - w_t\|_2$$

For $\Psi(w) = \tfrac{1}{2}\|w\|_2$, $D_\Psi(w'||w) = \tfrac{1}{2}\|w - w_t\|_2$

FTRL:

$$w_{t+1} = \arg\min_{w} \sum_{i=1}^{t} f(w, z_t) + \lambda_t \Psi(\mathrm{w})$$

Linearized FTRL:

$$w_{t+1} = \arg\min_{w} \left\langle \frac{1}{m}\sum_{i=1}^{t} \nabla f(w_t, z_t)\,, w\right\rangle + \lambda_t \Psi(w)$$

$\equiv$ (Stochastci) Mirror Descent:

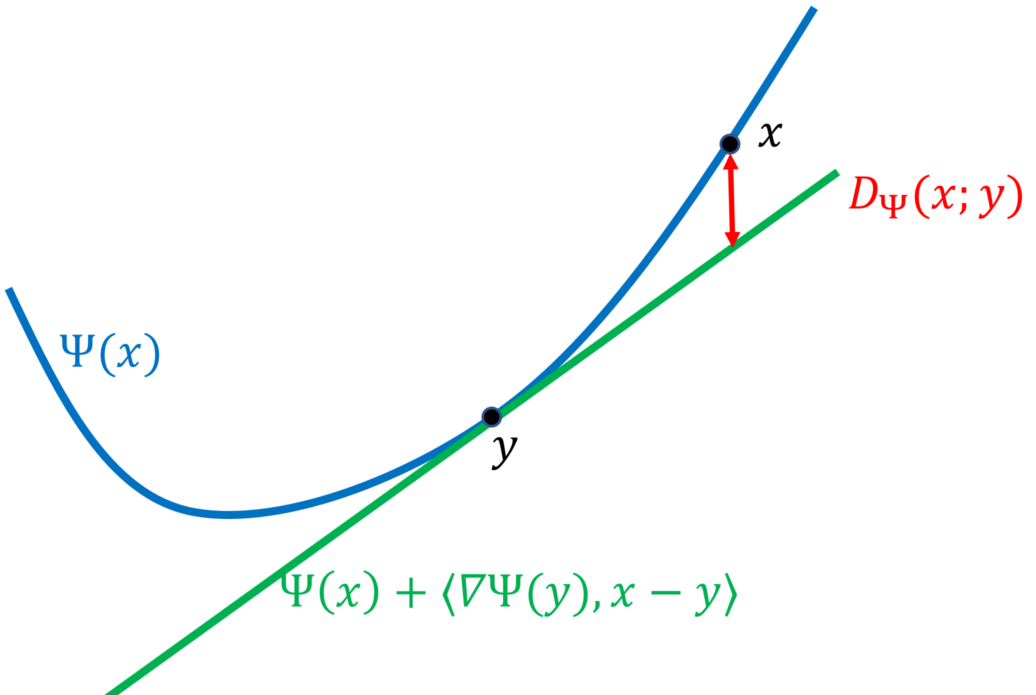
$$= \arg\min_{w\in\mathcal{W}} \langle \nabla f(w_t, z_t), w\rangle + \lambda_t D_\Psi(w||w_t)$$

Bergman Divergence: $D_\Psi(w'||w) = \Psi(w') - (\Psi(w) + \langle \nabla\Psi(w), w' - w\rangle)$

Bergman Divergence

$$D_\Psi(x;y) = \Psi(x) - (\Psi(y) + \langle \nabla\Psi(y), x-y\rangle)$$

- Ψ convex $\Leftrightarrow D_\Psi(x;y) \geq 0$
- Ψ strictly convex $\Rightarrow D_\Psi(x;y) = 0$ only for $x = y$
- Ψ α-strongly convex w.r.t. $\|x\|$ $\Rightarrow D_\Psi(x;y) \geq \frac{\alpha}{2}\|x-y\|^2$



Mirror Descent

$D_\Psi(w'||w) = \Psi(w') - (\Psi(w) + \langle \nabla\Psi(w), w' - w\rangle)$

$\Pi_\Psi^{\mathcal{W}}(w) = \min_{w' \in \mathcal{W}} D_\Psi(w'||w)$

$$w_{t+1} = \arg\min_{w \in \mathcal{W}} \langle \nabla f(w_t, z_t), w\rangle + \lambda_t D_\Psi(w||w_t)$$

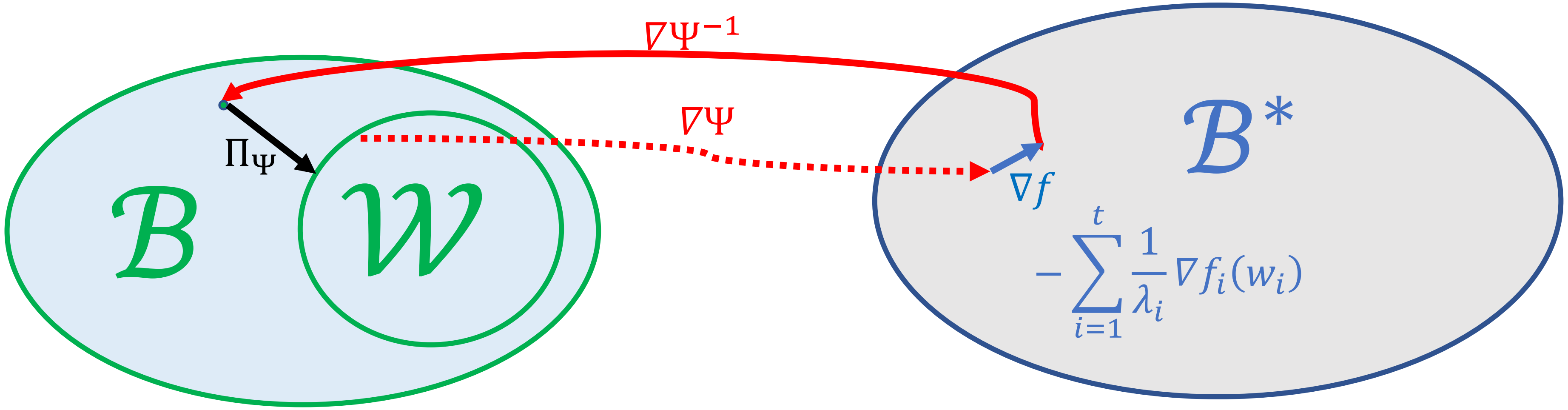
$$= \Pi_\Psi^{\mathcal{W}}\left(\nabla\Psi^{-1}\left(\nabla\Psi(w_t) - \frac{1}{\lambda_t}\nabla f(w_t, z_t)\right)\right)$$

If $\mathcal{W} = \mathcal{B}$
(no projections)

$$= \nabla\Psi^{-1}\left(\nabla\Psi(w_0) - \sum_{i=1}^t \frac{1}{\lambda_i}\nabla f(w_i, z_i)\right)$$

Init at
$w_0 = \arg\min \Psi(w)$

$$= \arg\min_{w} \sum_{i=1}^t \frac{1}{\lambda_i}\langle \nabla f(w_i, z_i), w\rangle + \Psi(w)$$



Optimization with Geometry Ψ

- Bergman Divergence: $D_\Psi(w'||w) = \Psi(w') - (\Psi(w) + \langle \nabla\Psi(w), w' - w\rangle)$

$$w^{(k+1)} = \arg\min_w \left\langle \nabla F\left(w^{(k)}\right), w\right\rangle + \frac{1}{2\eta} D_\Psi\left(w^{(k)}||w\right)$$

Mirror Descent

$$\approx \arg\min_w \left\langle \nabla F\left(w^{(k)}\right), w\right\rangle + \frac{1}{2\eta}\left(w - w^{(k)}\right)^\top \nabla^2\Psi\left(w^{(k)}\right)\left(w - w^{(k)}\right)$$

$$= w^{(t)} - \eta\left(\nabla^2\Psi\left(w^{(k)}\right)\right)^{-1} \nabla F\left(w^{(k)}\right)$$

Natural Gradient Descent

- Taking $w(\eta k) = w^{(k)}, \eta \to 0$:

$$\dot{w}(t) = -\nabla^2\Psi\left(w(t)\right)^{-1}\nabla F\left(w(t)\right)$$

Gradient Flow w.r.t
$\rho(w) = \nabla^2\Psi(w)$

- Discretizing corresponds to:

$$\dot{w}(t) = -\nabla^2\Psi\left(w([t]_\eta)\right)^{-1} \nabla F\left(w([t]_\eta)\right)$$

Natural Gradient Descent

$$\dot{w}(t) = -\nabla^2\Psi\left(w(t)\right)^{-1}\nabla F\left(w([t]_\eta)\right)$$

Mirror Descent

where $[t]_\eta = \eta\lfloor {}^t/_\eta \rfloor$

Optimization with Geometry Ψ

- Bergman Divergence: $D_\Psi(w'||w) = \Psi(w') - (\Psi(w) + \langle \nabla\Psi(w), w' - w\rangle)$

$$w^{(t+1)} = \arg\min_w \langle \nabla f(w^{(k)}, z_t), w\rangle + \frac{1}{2\eta_t} D_\Psi(w^{(k)}||w)$$

Stochastic MD

$$\approx \arg\min_w \langle \nabla f(w^{(k)}, z_t), w\rangle + \frac{1}{2\eta_t}\left(w - w^{(k)}\right)^\top \nabla^2\Psi\left(w^{(k)}\right)\left(w - w^{(k)}\right)$$

$$= w^{(k)} - \eta_t \left(\nabla^2\Psi\left(w^{(k)}\right)\right)^{-1} \nabla f(w^{(k)}, z_t)$$

Stochastic NGD

- Taking $w(\eta k) = w^{(k)}, \eta \to 0$:

$$\dot{w}(t) = -\nabla^2\Psi\left(w(t)\right)^{-1} \nabla F\left(w(t)\right)$$

Gradient Flow on **Population** w.r.t Ψ

- Discretizing linear approx. AND stochasticity:

$$\dot{w}(t) = -\nabla^2\Psi\left(w([t]_\eta)\right)^{-1} \nabla f\left(w([t]_\eta), z_{[t]_\eta}\right)$$

Stochastic NGD

$$\dot{w}(t) = -\nabla^2\Psi\left(w(t)\right)^{-1} \nabla f\left(w([t]_\eta), z_{[t]_\eta}\right)$$

Stochastic MD

where $[t]_\eta = \eta\lfloor t/\eta \rfloor$

Beyond the Euclidean Geometry

- SAA/(R)ERM Learning (Explicit Regularization):

$$\hat{w}_\lambda = \arg\min F_S(w) + \lambda\Psi(w)$$

$$\hat{w}_B = \arg\min L_S(w) \; s.t. \; \Psi(w) \le B$$

- SA Approach: "*Stochastic Mirror Descent*"

$$w^{(t+1)} = \arg\min_w \left\langle \nabla f\left(w^{(t)}, z_t\right), w \right\rangle + \eta_t D_\Psi\left(w^{(t)} || w\right)$$

- If $\Psi(w)$ is 1-strongly convex w.r.t. $||w||$:

$$L\left(\bar{w}^{(m)}\right), L(\hat{w}_\lambda), L(\hat{w}_B) \le L(w^*) + O\left(\sqrt{\frac{\Psi(w^*)\,\|\nabla f\|_*^2}{m}}\right)$$

$\|\phi(x)\|_*^2$ for
$f(w,x,y) = \ell(\langle w, \phi(x)\rangle, y)$
$|\ell'| \le 1$

For any convex
$F(w) = \mathbb{E}[f(w,z)]$,
by stability

Only for $F(w) =$
$L(w) = \mathbb{E}[\ell(\langle w, x\rangle, y)]$,
by uniform convergence

Matching the Geometry

For SMD discretization, $L\left(\bar{w}^{(m)}\right) \leq L(w^*) + O\left(\sqrt{\frac{\Psi(w^*)\ \|\nabla f\|_*^2}{m}}\right)$ if Ψ 1-s.c. w.r.t $\|w\|$

➔ need $m \propto \Psi(w^*)\ \|\nabla f\|_*^2 = \Psi(w^*)\ \|x\|_*^2$

- $\Psi(w) = \frac{1}{2}\|w\|_2^2$ is 1-strongly convex w.r.t $\|w\|_2$

$$\dot{\boldsymbol{w}} = -\boldsymbol{\nabla F}(\boldsymbol{w})$$
$$m \propto \|w\|_2^2 \cdot \|x\|_2^2$$

- $\Psi(w) = \frac{1}{2} w^T Q w$ is 1-strongly convex w.r.t $\|w\|_Q = \sqrt{w^T Q w}$

$$\dot{\boldsymbol{w}} = -\boldsymbol{Q}^{-1}\boldsymbol{\nabla F}(\boldsymbol{w})$$
$$m \propto (w^T Q w)(x^T Q^{-1} x)$$

- $\Psi(w) = \sum_i w[i] \log \frac{w[i]}{1/d}$ is 1-strongly convex w.r.t $\|w\|_1$ on $\{w > 0 | \|w\|_1 \leq 1\}$

$$\dot{\boldsymbol{w}}[\boldsymbol{i}] = -\boldsymbol{w}[\boldsymbol{i}]\boldsymbol{\partial_i F}(\boldsymbol{w})$$
$$m \propto \|w\|_1^2 \|x\|_\infty^2 \log(d)$$

- For smooth objectives wrt $\|\cdot\|$,

$$f(w',z) \le f(w) + \langle \nabla f(w,z), w' - w\rangle + \frac{H}{2}\|\Delta w\|^2$$

- Or, for "relative smooth" objectives:

$$f(w',z) \le f(w) + \langle \nabla f(w,z), w' - w\rangle + H D_\Psi(w'||w)$$

for differentiable f, Ψ equivalent to:

$$\nabla^2 F(w) \preccurlyeq H \nabla^2 \Psi(w)$$

$$\begin{aligned}\mathbb{E}[F(\bar{w}_{\mathrm{SMD}})] - F(w^*) &\le O\left(\frac{H\Psi(w^*)}{T} + \sqrt{\frac{\mathbb{E}\left[\|\nabla f(w^*,z) - \nabla F(w)\|_*^2\right] \cdot \Psi(w^*)}{T}}\right) \\ &\le O\left(\frac{H\Psi(w^*)}{T} + \sqrt{\frac{HF(w^*)\Psi(w^*)}{T}}\right) \qquad (?)\end{aligned}$$

For $f(w,z) \ge 0$

$$\dot{w}(t) = -\rho\big(w(t)\big)^{-1}\nabla F\big(w(t)\big)$$

- Natural Gradient Descent:

$$\dot{w}(t) = -\rho\left(w([t]_\eta)\right)^{-1}\nabla f\left(w([t]_\eta), z_{[t]_\eta}\right)$$

$$\Rightarrow\ w_{k+1} = w_k - \eta\rho\big(w(t)\big)^{-1}\nabla f(w_k, z_k)$$

- Mirror Descent:

$$\dot{w}(t) = -\rho\big(w(t)\big)^{-1}\nabla f\left(w([t]_\eta), z_{[t]_\eta}\right)$$

$\Rightarrow$ w_{k+1} is obtained from solution to

$$\dot{w}(t) = -\rho\big(w(t)\big)^{-1} g_k \qquad g_k = \nabla f(w_k, z_k)$$

Steepest Descent

Steepest descent w.r.t. a $\delta(w', w)$ (perhaps not even a divergence):

$$w_{t+1} = \arg\min_{w} \langle \nabla f(w_t, z_t), w \rangle + \lambda_t \delta(w_t, w)$$

✓improve the objective as much as possible

✓only a small change in the model.

Examples:

- Steepest descent w.r.t $\delta(w', w) = \|w' - w\|_2$ ➔ Gradient Descent
- $\delta(w', w) = \|w' - w\|_1$ ➔ coordinate descent
- $\delta(w', w) = \|w' - w\|_\infty$ ➔ $\Delta w \propto sign(\nabla f)$

- So far: Implicit regularization of one-pass SGD
  - Important that we use fresh example at each iteration
  - Only one pass over the data
  - Number of opt iterations = Number of data points
  - We do not get to zero training error (even if it’s possible)
  - In a sense: regularization from early stopping

- Neural Net training phenomena:
  - Many passes of SGD
  - Optimize to zero training error
  - No early stopping